INTRODUCTORY SCIENCE TEXTS

Low Noise Electronics

QUANTUM ELECTRONICS
MASER . LASER
PLASMA WAVES
PARAMETRIC AMPLIFIERS
NOISE TEMPERATURE
TRAVELLING WAVE TUBE
ADLER TUBE . PHONON

W.P. JOLLY

Low Noise Electronics W.P.JOLLY

LOW NOISE ELECTRONICS

This book is one of a series of short, simple, treatments of advanced modern subjects. It has been written so that the whole range of the topic may be grasped quickly by a reader with the general scientific background of a first or second year college student. The text is supported by references to books and journals carefully chosen to extend any section well into the post-graduate field.

The book should be valuable to working graduates who wish to keep up to date, particularly science teachers, to students of electrical engineering, and to science specialists in colleges of education.

25s 0d net
IN UK

LOW NOISE ELECTRONICS

An introduction to quantum and electron beam electronics

Introductory Science Texts

Editor: Professor W. P. Jolly

LOW NOISE ELECTRONICS

W. P. Jolly, B.Sc., C.Eng., F.I.E.E.
Assistant Professor in the Dept. of Physics and Electrical Engineering
Royal Naval College, Greenwich.

HIGH FREQUENCY COMMUNICATIONS

J. A. Betts, B.Sc., Ph.D., C.Eng., M.I.E.E.
Department of Electronics
University of Southampton.

In preparation:

HYBRID COMPUTATION
CRYOELECTRONICS

Low Noise Electronics

W. P. JOLLY, B.Sc., C.Eng., F.I.E.E.

Assistant Professor in the Department of Physics
and Electrical Engineering
Royal Naval College, Greenwich

THE ENGLISH UNIVERSITIES PRESS LTD
St Paul's House Warwick Lane London EC4

First printed 1967

Printed and bound in Great Britain for
THE ENGLISH UNIVERSITIES PRESS LIMITED
by Richard Clay (The Chaucer Press), Ltd.,
Bungay, Suffolk

Editor's Foreword

The very rapid growth of engineering and the physical sciences in recent years has meant that many new subjects have appeared which are so important fundamentally and technologically that they could well be studied in an undergraduate course. The degree and diploma syllabuses are, however, already so crowded that little time can be found for new subjects and a post-graduate year of extra tuition is generally proposed to meet this difficulty.

There are many students and busy practising scientists and engineers who will wish to have a working knowledge of important and growing new subjects but who have no time to study them at length.

The books in this series are intended to be short and simply written so that they may be read in a few days by anyone with the general background of a second-year student.

The text is supported by references to published papers and textbooks, chosen from the large mass of available material as being most likely to help the reader who may wish to pursue the subject, or some part of it, in greater depth now or later.

W. P. JOLLY

Author's Preface

I have tried to produce a book which really is an introduction to the subject. Masers, lasers, electron beam tubes, parametric amplifiers and associated topics in low noise and quantum electronics have been treated simply, and any supporting material needed, e.g. semiconduction, has been developed from fundamentals.

Sometimes the full and detailed discussion of a principle has been thought too long or difficult for a book which aims to be a primer. In such cases the principle is stated and discussed simply in the text, and reference made to an original paper, review article, or advanced textbook where a full account may be found.

The bibliography lists a number of books for the reader who wishes to extend his knowledge, and suggests journals which should be consulted from time to time to keep informed of new developments.

There are no toy cricket bats at Lords but many children get the feel of the game by using a bat light enough to handle with ease. I would like to feel that this book filled a similar role.

W. P. JOLLY

Contents

1

Continuous Spectrum Radiation

SOURCES OF ELECTROMAGNETIC RADIATION usually generate energy either over a wide continuous band of frequencies, or at a comparatively small number of discrete frequencies. Typically, a hot solid is a source of the first kind and an electrically excited gas in a discharge tube is a source of the second kind.

The study of the continuous spectrum of radiation from a hot solid led Planck, in 1900, to formulate the quantum theory. This chapter will be concerned with the quantum theory of radiation and its special application in modern communications. The next chapter will show how Bohr applied the quantum theory to explain the line spectra of the radiation emitted by sources of the second kind.

Let us first review some of the ideas about continuous spectrum radiation which were already established before the quantum theory.

Pre-quantum Ideas

By the end of the nineteenth century a number of important qualitative concepts about radiation had become accepted.

It was known that good absorbers were good radiators, and that the best radiator was therefore a black body. This was defined as a body which absorbed completely all the radiation falling on it. Also the theory of exchanges had shown that the transfer of energy by radiation from a hot to a cold body was a nett process—a hot body radiating energy to a cold one was at the same time receiving energy from it.

A constant temperature enclosure containing various bodies will be filled with radiation as energy passes to and fro between the bodies and the walls. Each body radiates energy of exactly the same amount and character as it absorbs, otherwise it would change its temperature. Thus the energy density of the radiation in the enclosure, and its spectral distribution, depend only on the temperature.

A black body in the enclosure will absorb all the energy falling on it and will thus radiate the same energy. Thus the amount and spectral

character of the radiation emitted by a black body is exactly the same as that which falls on the body in an enclosure at the same temperature.

In addition to qualitative ideas of this type there were two important pre-quantum laws of a quantitative kind. These are Kirchhoff's and Stefan's laws which are dealt with below.

Kirchhoff's Law

Kirchhoff's law states the connection between $e_\lambda \, . \, d\lambda$, the emissive power, and the absorptive power of a body at a given temperature.

If a body emits $e_\lambda \, . \, d\lambda$ of energy per unit area per second in the form of radiation with wavelength lying between λ and $\lambda + d\lambda$, then e_λ is called the emissive power.

If a given amount of radiation in this same waveband falls on a body and a fraction a_λ is absorbed, then a_λ is called the absorptive power.

Consider unit area of any body in a constant temperature enclosure. The radiation falling on it is dQ, and the amount absorbed $a_\lambda \, . \, dQ$ equals the radiation emitted by it and depends only on the temperature.

Thus $$e_\lambda \, . \, d\lambda = a_\lambda \, . \, dQ$$

$$\frac{e_\lambda}{a_\lambda} = \frac{dQ}{d\lambda}$$

which is constant for all bodies at a given temperature.

Stefan's Law

Stefan's law is concerned with the total power e of all wavelengths emitted by unit area of a black body at temperature T.

$$e = \sigma T^4$$

where σ is Stefan's constant $5{\cdot}735 \times 10^{-8}$ watt/m^2/deg^4.

With Stefan's law established for the total amount of radiation of all wavelengths emitted by a black body it only remained to show how this energy was distributed among the various wavelengths.

Distribution Laws. Planck's Formula

If a satisfactory explanation were to be provided for the spectrum of radiation contained in a constant temperature enclosure, or emitted by a black body, then a plausible statement had first to be made about the number and behaviour of the individual sources of energy at the various frequencies. Statistical mechanics and classical wave theory provided such a statement.

Statistical mechanics is concerned with the use of the laws of mechanics and of probability to predict the equilibrium state of a system. Such a treatment shows that, of a large number of oscillators of a given frequency, the number possessing energy between E and $E + dE$ is equal to $Ae^{-\frac{E}{kT}} . \, dE$, where A is a constant depending on the total number of oscillators present, T is the absolute temperature, and k is Boltzmann's constant. This is Boltzmann's law and is fully treated in Jean's *Dynamic Theory of Gases*.

Planck introduced his novel quantum concept and used it in the Boltzmann expression to obtain the mean energy of all the oscillations in the frequency range f to $f + df$. He suggested that any oscillator at frequency f could only lose or gain energy in multiples of hf, the single quantum. Thus the values which E can take in the Boltzmann expression are 0, hf, $2hf$, $3hf$, etc.

The number of oscillations with these energies will be N, $Ne^{-hf/kT}$, $Ne^{-2hf/kT}$, etc., where N is the number present in the lowest energy state.

Summing all these terms gives the total number of oscillations present as

$$\frac{N}{1 - e^{-\frac{hf}{kT}}} \quad . \quad . \quad . \quad . \quad . \quad (a)$$

Total energy present is $N(hfe^{-hf/kT} + 2hfe^{-2hf/kT} + \ldots)$

$$= \frac{Nhf}{e^{\frac{hf}{kT}}\left(1 - e^{-\frac{hf}{kT}}\right)^2} \quad . \quad . \quad . \quad . \quad . \quad (b)$$

Classical wave theory may now be used to calculate the number of oscillations of frequency between f and $f + df$ in unit volume of an enclosure. This figure is then equated to (a) above, so that the unknown N may be eliminated from (b) which gives the total energy at frequency f.

The number of oscillations per unit volume is found to be $\frac{8\pi}{\lambda^4} d\lambda$ or $\frac{8\pi f^2}{c^3} df$, where c is velocity of light.[1]

Thus $$\frac{N}{1 - e^{-hf/kT}} = \frac{8\pi f^2}{c^3} df$$

$$N = (1 - e^{-hf/kT}) \frac{8\pi f^2}{c^3} df$$

[1] Roberts, *Heat and Thermodynamics* (1940), p. 416.

and substituting in (*b*) gives for the total energy per unit volume in the frequency range f to $f + df$

$$E_f \,.\, df = \frac{8\pi f^2}{c^3} \frac{hf}{e^{hf/kT}\,(1 - e^{-hf/kT})}\, df$$

$$= \frac{8\pi f^2}{c^3}\left(\frac{hf}{e^{hf/kT} - 1}\right) df$$

which is Planck's distribution law.

If $E_f \,.\, df$ is the energy density, then the energy falling per second on unit area in the enclosure is $E_f \,.\, df \times \frac{c}{4}$, and this is the power radiated by unit area of a black body.[2]

Note that dividing (*b*) by (*a*) gives for the mean energy of each oscillation

$$\left(\frac{hf}{e^{hf/kT} - 1}\right)$$

instead of the classical equipartition value kT.

Comparison of Quantum and Classical Views of Radiation

The difference between Planck's law and earlier, classical distribution laws is the substitution for kT of the frequency dependent expression

$$\frac{hf}{e^{hf/kT} - 1}$$

as the energy associated with an oscillator.

$e^{hf/kT}$ may be expanded as $1 + \frac{hf}{kT} + \frac{1}{2}\left(\frac{hf}{kT}\right)^2$ and if $hf \ll kT$, then the quantum expression reduces to the classical value kT and Planck's law reduces to the Rayleigh–Jeans expression

$$E_f \,.\, df = \frac{8\pi f^2}{c^3}\, kT \,.\, df$$

The classical law fails if the frequency is high (the ultra-violet catastrophe), or if the temperature is low. The more accurate quantum expression must then be used.

Electronic Noise

The infinite frequency spectrum of radiation from a black body includes a significant power output in the "radio" frequency band. This radiation will be detected by a radio receiver and its presence will

[2] Roberts, *Heat and Thermodynamics* (1940), p. 388.

indicate that the black body is delivering power to the input terminals of the receiver either through the aerial or in some other fashion. This power will be amplified in the receiver and will appear at the output terminals, where it will in general act as a background against which the wanted signal must be detected.

The ability of the receiver to detect a signal is determined by the signal to noise power ratio, and the noise may arise from thermal radiators outside the receiver or from circuit elements within it. In particular, a resistance across the input terminals of a receiver will act as a noise source.

Nyquist[3] in 1928 considered the particular case of a resistance as a noise source and derived the expression kTB for the available noise power in the bandwidth B. The Nyquist expression, which is of considerable generality, was originally derived in classical equipartition terms and the term kT should more properly be replaced by the quantum term.

But just as Planck's accurate quantum distribution law reduces to the classical Rayleigh–Jeans expression when $hf \ll kT$, so the kTB expression is of sufficient accuracy in many radio fields.

It is instructive to compare the "classical" and the "correct" noise powers in the microwave region, at room temperature and at a typical cryogenic temperature.

For ease of calculation let us take 30,000 Mc/s, i.e. $\lambda = 1$ cm, as a typical microwave frequency.

We are concerned with the discrepancy between kT and

$$\frac{hf}{e^{hf/kT} - 1}$$

i.e. between $\frac{hf}{kT}$ and $\frac{ht}{kT}\left(1 + \frac{1}{2}\frac{hf}{kT}\right)$ if only the first two terms of the expansion are considered.

The comparison between 1 and $\left(1 + \frac{hf}{2kT}\right)$ gives an indication of the discrepancy.

For 30,000 Mc/s at 300° K $\frac{hf}{2kT} = \frac{3 \times 10^{10} \times 6{\cdot}39 \times 10^{-34}}{2 \times 300 \times 1{\cdot}38 \times 10^{-23}} < 3 \times 10^{-3}$

For 30,000 Mc/s at 1° K $\frac{hf}{2kT} = 0{\cdot}69$

Thus the use of the classical expression kTB to calculate noise power in the microwave band at room temperature leads to little error.

[3] Nyquist, *Phys. Rev.*, Vol. 32, 1928, p. 110.

But at 1° K the calculated noise power using kTB may be very much greater ($\simeq$ 50% greater) than that actually experienced.

Noise Input to Directional Aerial from Large Distant Source

As an example of the generality of Nyquist let us consider an aerial looking out into space.

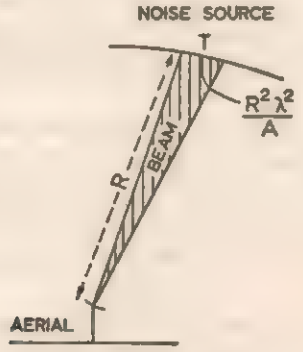

FIG. 1.1

The directional aerial has dimensions comparable with the wavelength of the radiation being received, and it is in a narrow band centred on this wavelength that the noise input is of interest.

Planck's distribution law for $E_f \cdot df$, the energy density in an enclosure at frequency f, is

$$E_f \cdot df = \frac{8\pi}{c\lambda^3}\left(\frac{hf}{e^{hf/kT}-1}\right)df$$

Power radiated by unit area of black body in this band is

$$\frac{E_f c}{4} \cdot df = \frac{2\pi}{\lambda^3}\left(\frac{hf}{e^{hf/kT}-1}\right)df$$

(see p. 4)

Diffraction determines the beam width of the aerial as $\frac{\lambda^2}{A}$, where A is area of aerial.[4]

Beam width and range R determine the area of the noise source which radiates to aerial as $\frac{R^2\lambda^2}{A}$.

Power radiated towards aerial is

$$P = \frac{R^2\lambda^2}{A} \cdot \frac{2\pi}{\lambda^3}\left(\frac{hf}{e^{hf/kT}-1}\right)df$$

[4] *Physics for Electrical Engineers*, Jolly, E.U.P. (1961).

Power per unit area at aerial is

$$\frac{P}{2\pi R^2}$$

Thus power received by aerial is

$$\frac{PA}{2\pi R^2} = \frac{R^2\lambda^2}{A} \cdot \frac{2\pi}{\lambda^3} \cdot \frac{A}{2\pi R^2}\left(\frac{hf}{e^{hf/kT}-1}\right)df$$

$$= \left(\frac{hf}{e^{hf/kT}-1}\right)df = \left(\frac{hf}{e^{hf/kT}-1}\right)B$$

if bandwidth B is small. With the usual reservations this reduces to kTB.

Noise Temperature

Commonly a radiator at temperature T will not be a black body but will radiate energy as though it were a black body at a different temperature T'. T' is called the colour temperature of the body in pyrometry and the noise temperature in electronics.

Receiver Noise, Noise Input Temperature

When a signal is amplified and processed by a receiver the ratio of signal to noise power at the output is less than that at the input because of the extra noise introduced by the various stages inside the receiver.

The noise figure of a receiver is defined as

$$F = \frac{S_i}{N_i}\Big/\frac{S_o}{N_o}$$

where S and N are signal and noise power, and the subscripts refer to input and output.

Each stage of an amplifier will introduce some noise and the effect of this noise is greater if it is introduced by an early stage because it will then be amplified by the following stages.

It is convenient to express the effect of all the noise introduced by the various stages of the receiver in terms of the noise power N_{Ri}, which would have produced the same effect if introduced at the input terminals.

The power appearing at the output terminals due to noise introduced by the receiver is thus GN_{Ri}, where G is the power gain of the receiver.

Thus $$F = \frac{S_i}{N_i}\Big/\frac{S_o}{N_o} = \frac{N_o}{GN_i} = \frac{GN_i + GN_{Ri}}{GN_i} = \frac{N_i + N_{Ri}}{N_i}$$

i.e. $$\frac{N_{Ri}}{N_i} = F - 1$$

Note that it is N_{Ri} which is the fixed receiver property and that the noise figure F depends here upon N_i.

So that F shall also be a fixed receiver parameter, N_i is standardised as the noise from a source at 290° K.

Another way of considering the noise introduced by a receiver is to use the concept of receiver noise input temperature.

The input noise temperature T_i of a receiver is the increase in source temperature needed to double the output of the receiver when originally connected to a noise-free source.

Thus $N_{Ri} = kT_iB$

and $N_i = 290 \,.\, kB$

Whence $T_i = (F - 1) \times 290°\ \mathrm{K}$

A good conventional microwave receiver might have a noise figure of 10, giving it a noise input temperature of just over 2600° K.

Sometimes the noise figure is quoted in decibels, in which case $T_i = 290\ (10^{x/10} - 1)$, where x is the noise factor in db.

The new unconventional amplifiers, e.g. masers, generate so little noise that their noise input temperature may be only a few degrees. In this field it is usual to work in noise temperature, but if a noise figure is quoted it will be very close to unity, e.g. $F = 1{\cdot}01$ gives 2·9° as input temperature.

The reader may possibly encounter the statement that a noise figure of 1·5 corresponds to a noise temperature of 2°. Such a statement will refer to some low temperature device operating in liquid helium at 4° K. This ambient temperature is used instead of the 290° of the formal definition. Such an alternative standard is confusing and should be avoided.

Noise Due to Lossy Medium

If e_λ is the emissive power of a real body and ${}_be_\lambda$ is that of a black body at the same temperature T, then using the classical approximation.

Power radiated by black body is ${}_be_\lambda \,.\, d\lambda = kTB$

Power radiated by real body is $e_\lambda \,.\, d\lambda = kT'B$

Thus $T'/T = e_\lambda / {}_be_\lambda$

But e_λ may be found in terms of a_λ the absorption coefficient at this temperature.

By Kirchhoff's law $e_\lambda/a_\lambda = {}_be_\lambda/{}_ba_\lambda = {}_be_\lambda$, where ${}_ba_\lambda = 1$ for black body.

Thus $e_\lambda = a_\lambda \,.\, {}_be_\lambda$

and $T'/T = a_\lambda \,.\, {}_be_\lambda/{}_be_\lambda$

$T' = a_\lambda T$

Any absorbing substance will act as a noise source with a noise temperature depending on its absorption coefficient and its real temperature. The lossy medium may be the atmosphere or an attenuating element in the feeder system, such as a lossy switch or waveguide.

A waveguide at room temperature giving 1% absorption would contribute about 3° K to the total noise temperature, which might be about 20° K in a typical low noise system. If the guide were cooled in liquid helium the contribution would be negligible (0·04° K).

Noise Temperature for Complete System

Published figures[5] for the noise performance of a high-quality system are worth studying in terms of some of the ideas developed earlier in this chapter.

Sky noise at zenith . .	2·5° K
Subsidiary lobes . . .	2° K
Aerial, joint, and feeder loss	12° K
Maser noise . . .	2° K
	18·5° K

The sky noise is quoted with the aerial pointing vertically upwards. This gives the minimum path through the absorbing atmosphere. At lower angles of elevation the path through the lossy atmosphere, which is at a relatively high temperature, will be longer and the noise temperature consequently higher.[6]

Any high-gain aerial may have small side or back lobes as well as the main beam. These may receive radiation from the ground and give subsidiary lobe contribution to the total noise temperature.

The feeders, switches, etc., are treated just as any lossy medium and

[5] "Overall System Requirements for Low Noise Performance", Ditchfield, *J. Brit. I.R.E.*, Vol. 22, August 1961.

[6] See also *Sky Noise*, p. 136.

will give noise depending on the attenuation they introduce and their actual temperature.

The contribution of the maser itself to the total noise temperature of the system[7] is very small compared with other types of amplifier and is discussed elsewhere.

[7] "System Noise Temperature of Quantum Amplifiers", Van der Ziel, *Proc. I.E.E.E.*, June 1963.

2

Radiation at Discrete Frequencies

THE NATURE OF THE continuous spectrum of electromagnetic radiation emitted by black bodies, and by solids which behave in a similar fashion, has already been considered, and Planck's quantum theory has emerged as a result of the attempts to explain the shape of the energy distribution curve.

In this section will be discussed those substances, principally gases, which only emit—or absorb—radiation of certain discrete wavelengths or within certain fairly narrow bandwidths. The subject stems from Bohr's use of Planck's quantum ideas to explain the line spectrum of hydrogen in terms of atomic energy levels.

Bohr Theory of Hydrogen Spectrum

The hydrogen atom contains only one electron and Bohr postulated that the electron rotated round the nucleus in a circular orbit.

Let the radius of the orbit be r and the electron velocity, mass, and charge v, m, and e.

Then, equating the centrifugal force and the Coulomb force of attraction

$$\frac{mv^2}{r} = \frac{e^2}{4\pi k_0 r^2}$$

Bohr further assumed that the angular momentum of the electron mvr could only be an integral multiple of $\frac{h}{2\pi}$, where h was Planck's constant.

$$mvr = \frac{nh}{2\pi}$$

where n is any integer.

Eliminating v from these two equations gives

$$r = \frac{n^2 k_0 h^2}{\pi e^2 m}$$

The total energy of the system is the kinetic energy of the moving electron plus the potential energy due to the separated positive and negative charges.

The potential energy of the electron at radius r is the potential energy at infinity E_∞, less the work performed as the electron comes in from infinity to r.

Potential energy is $E_\infty - \int_\infty^r \frac{e^2}{4\pi k_0 r^2}\,dr = E_\infty - \frac{e^2}{4\pi k_0 r}$

Kinetic energy is $\frac{1}{2}mv^2 = \frac{e^2}{8\pi k_0 r}$

Thus total energy is $E_\infty - \frac{e^2}{8\pi k_0 r}$ and $r = \frac{n^2 k_0 h^2}{\pi e^2 m}$

If the electron falls from an orbit of large radius r_2, characterised by quantum number n_2, to an orbit of smaller radius r_1 characterised by n_1, then energy $E_2 - E_1 = hf$ is radiated.

$$hf = \left(E_\infty - \frac{e^4 m}{8k_0^2 n_2^2 h^2}\right) - \left(E_\infty - \frac{e^4 m}{8k_0^2 n_1^2 h^2}\right) = \frac{e^4 m}{8k_0^2 h^2}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Since $c = f\lambda$, where c is the velocity of light,

$$\frac{1}{\lambda} = \frac{e^4 m}{8k_0^2 h^3 c}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

This expression, now deduced theoretically by Bohr, was of exactly the same form as that which had been formulated earlier by experimental spectroscopists to describe the wavelength of the lines of the hydrogen spectrum.

More Complex Atomic Models

The simple line spectrum of hydrogen can be explained in terms of Bohr's model. This model is based on classical mechanics and electrostatics with two innovations: the introduction of the Planck relation between energy and frequency $E = hf$, and the quantisation of angular momentum.

There remained a number of other spectral features, well established empirically but not explained by Bohr's simple model. Such features were fine structure, hyperfine structure, and electro- or magneto-optical effects, such as the Stark and Zeeman effects, where the application of an electric or magnetic field to the light beam before dispersion in the spectrometer produces splitting of the spectral lines.

In order to account for these more detailed spectral properties the simple model was elaborated by the addition of such concepts as elliptical orbits, electron spin, and nuclear spin. The extra properties introduced were quantised, just as angular momentum had been in the simple model. As a result the widely separated single electronic energy levels of the original model were replaced by sets of closely spaced levels which accounted for the fine and hyperfine structure of the spectral lines.

Because of Planck's relation $E = hf$ between energy and frequency in a quantum process, it will be necessary for a transition to occur between a pair of very closely spaced levels if an atom is to radiate or absorb energy in the microwave region.

Such a transition occurs in the hydrogen atom at 1420 Mc/s, and has been detected in interstellar radiation. It is associated with the magnetic interaction of nuclear and electron spin, and the energy change involved when the electron changes its spin from one direction to the other.

The caesium atom has a single electron in its outermost orbit. The spin of this electron reacts with the nuclear spin in a similar fashion to that of the hydrogen atom to produce a transition, in the case of caesium, at 9192 Mc/s. This transition is used to produce the frequency standard in the caesium atomic clock. The frequency standard is an oscillator which supplies energy to the stream of caesium atoms. When the oscillator is at precisely the transition frequency a maximum amount of power is absorbed. A control system monitors the amount of power absorbed and alters the oscillator frequency until maximum absorption is obtained. Frequency stability better than one part in 10^9 has been obtained.

Molecular Structure

The energy-level picture for a molecule will be correspondingly more complex than that for an atom, because there will be at least two atomic systems in close proximity. These will each react upon the other to produce increased complexity in the original system. Further energy levels are introduced because the molecule may possess vibrational or rotational energy.

So complex is the energy-level structure that the molecular spectra tend to consist of bands of extremely closely spaced levels. Microwave absorption spectroscopy is a useful way of determining the detailed energy-level structure.

Although it is often possible to discuss some particular feature of a molecular spectrum in terms of a model, it is only a quantum mechanical solution of the problem which reveals the full details of the energy

levels and justification of the selection rules which state whether transitions between certain levels are permitted or not.

The ammonia molecule was the first in which microwave effects were observed and its study has consistently yielded results important in the microwave field. Its inversion resonance at 24 Gc/s was the basis of the first maser and can be explained in model terms because there are two stable structures for the NH_3 molecule. One structure has the nitrogen atom above the plane of the three hydrogen atoms, and the other has the nitrogen atom below the plane. A transition between the two states involves a quantum of energy at 24 Gc/s.

Oxygen and Water Molecules

Oxygen and water molecules are of particular interest because they occur naturally in the Earth's atmosphere, and because their complex molecular spectra contain closely spaced energy levels between which transitions may occur with the absorption of energy in the microwave frequency region.

Electrically non-polar molecules like oxygen and nitrogen normally only have resonances at ultra-violet frequencies, corresponding to transitions between two electronic states. But oxygen is paramagnetic[1] and, because of the permanent magnetic moment of the oxygen molecule, transitions are permitted by the selection rules between closely spaced components of the ground state which give rise to microwave resonances.

The microwave spectrum of the oxygen molecule consists of an isolated line at 2·5 mm (120 Gc/s) and a set of about twenty-five major lines at around 5 mm (60 Gc/s) which are so close together that they are unresolved at atmospheric pressure (see below).

The water molecule[2] has many pairs of nearly coincident energy levels which are close enough together for the absorption to lie in the microwave spectrum. In practically all cases, however, the selection rules forbid transitions between these levels, and there is only one strong absorption line in the microwave region, at 1·35 cm (22 Gc/s). The remainder of the permitted transitions give resonance peaks in the millimetre and in the infra-red regions.

Atmospheric Absorption

The general character of the variation with frequency of atmospheric absorption in the microwave region is shown in Fig. 2.1. A more

[1] "Absorption of Microwaves by Oxygen", Van Vleck, *Phys. Rev.*, Vol. 71, No. 7, p. 413.

[2] "Absorption of Microwaves by Uncondensed Water Vapour", Van Vleck, ibid., p. 425.

quantitative graph showing the contribution of atmospheric absorption to the total noise temperature of the sky is given elsewhere (p. 137). The purpose of this section is merely to indicate why the atmospheric absorption spectrum is continuous—although peaky—from 1 Gc/s upwards, in spite of the fact that the microwave spectra of oxygen and water vapour consist of discrete lines at 22 Gc/s and certain other higher frequencies.

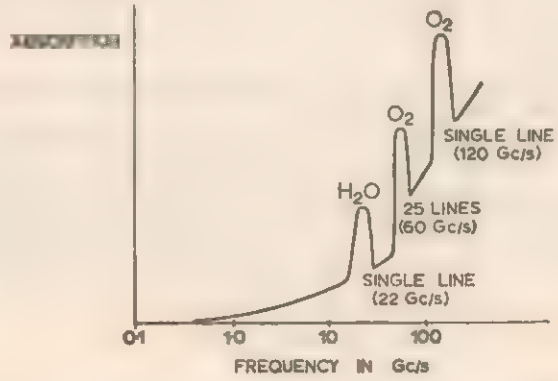

FIG. 2.1

There are three types of mechanism which contribute to the frequency broadening of a single spectral line.[3] The three types, together with the order of magnitude of the broadening produced in a microwave resonance, are: natural width (<1 cycle), Doppler effect (50 kc/s), and collision broadening (several Gc/s at atmospheric pressure).

The natural width of a line increases with the spontaneous emission probability, rather as the Q of an oscillating circuit is lowered by loading. The width increases with frequency and is significant at X-ray frequencies but negligible for microwaves.

The Doppler broadening is due to the relative motion of the molecule and the observer. It depends on the kind of molecule and its temperature, but is typically 50 kc/s, which is not important in the communications field, although it may be important in microwave frequency standards like the caesium clock.

The discrete frequencies in the absorption and emission spectrum of a given gas are calculated for a single isolated molecule. If another molecule were close enough to interact with the first, then the spectrum would be modified. Collision broadening is the summation of the frequency broadening effects caused by interaction of molecules which are

[3] *Microwave Spectroscopy*, Gordy *et al.*, Chapman & Hall.

instantaneously close together due to random motion in a natural sample of gas.

The magnitude of the collision-broadening effect depends, among other things, upon the nature of the colliding gas molecules and upon the pressure. It is, however, large compared with the other two effects in the microwave region, and causes the individual oxygen and water-vapour lines to run into each other in the microwave spectrum as shown in Fig. 2.1. A significant part of the absorption curve shown is due to transitions of high absorption in the water molecule which occur in the millimetre and infra-red regions beyond 120 Gc/s. Some of these are pressure broadened so that the skirts of the lines contribute to the absorption at much lower frequencies.

The theoretical calculation of the strength and shape of an absorption line in a given molecule and the experimental results[4] show good agreement. The extrapolation of the theory to explain the absorption by the atmosphere, with all the uncertainties of its composition, is difficult. Experiments produce unexpected results, such as the suggestion[5] that absorption is much higher than would be expected when oxygen and water-vapour lines fall close together.

Other Line Spectra

The particular gases discussed above do not make up an exhaustive list of those with interesting and important microwave properties. Although they are the most important in the communication field at present (see also, p. 137), there is no reason why other molecules should not be important in the future. For instance, recent reports[6] describe the detection of the absorption in interstellar space of radiation at 1660 Mc/s. This is one of the lines in the microwave spectrum of hydroxyl (OH) which is thus assumed to exist in space.

In discussing pressure broadening above we have touched upon the effects of interaction between closely spaced molecules. These effects are accentuated in the solid state (p. 21), although isolated impurity atoms in an inactive host material (e.g. Cr in Al_2O_3) may still give well-defined microwave lines of the type described above.

[4] Schulze and Tolbert, *Nature*, Vol. 200, November 1963, p. 747.
[5] Tolbert and Straiton, *P.I.R.E.*, March 1961, p. 649.
[6] "Radio Observations of OH in the Interstellar Medium", Weinreb *et al.*, *Nature*, Vol. 200, No. 4909, November 1963, p. 829, and ibid., Vol. 208, No. 5009 October 1965, p. 640.

3

Quantum Picture of Electrons in Solids

WHEN MANY ATOMS ARE packed close together in the solid form most of the interesting electrical properties, e.g. conductivity and the way in which the material reacts to electromagnetic radiation at optical and microwave frequencies, can be explained in terms of the behaviour of certain of the electrons in the material. In particular, the energy possessed by the electrons and the way in which it may change is important.

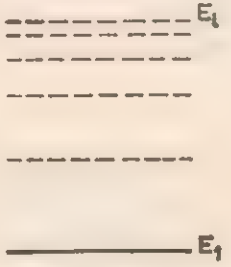

FIG. 3.1

The total energy possessed by the electrons in a solid element is distributed among them in a complex fashion, but the elaborate electronic energy distribution in the solid is evolved from the more simple energy distribution of each constituent atom—the characteristic energy-level structure of an isolated atom of the particular element.

We have already developed a quantitative theory for the electron energy levels in the isolated hydrogen atom. It is instructive to consider the extent to which the general principles established about the energy-level structure of hydrogen and other multi-electron atoms can be extended into the solid state.

The permitted energy levels for the electron in the hydrogen atom have the general form shown in Fig. 3.1 E_1 is the ground or normal state for the atom and E_t is the maximum energy the electron may

possess without ionisation occurring. The full line for E_1 indicates that the particular atom shown is in the ground state, i.e. that this energy level is occupied by an electron. The dotted lines for the higher energy levels indicate that in this case these permitted levels are vacant.

If the electron possesses more energy than E_4, then it is no longer confined to one of the permitted orbits near the nucleus but may wander farther away without restriction.

When models of the atoms containing more than one electron were constructed and explanations produced for the spectra of these heavier elements, another important principle was recognised which will be

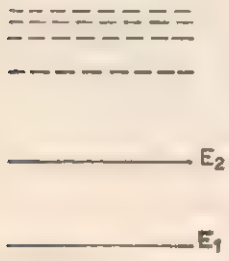

FIG. 3.2

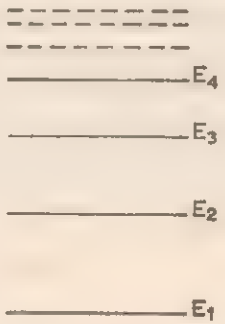

FIG. 3.3

extended into the more complex system of the multi-atom solid crystal. This principle was Pauli's exclusion principle, which stated that no more than two electrons in an atom could possess the same energy, and these two electrons must have opposite spins.

Thus the helium atom in the ground state has two electrons of opposite spin in a single orbit. Lithium, with three electrons, has two of opposite spins in its inner orbit, which is thus completely filled, and the third electron is in an orbit of greater radius.

The energy-level picture for the electrons in the isolated lithium atom is of the general form shown in Fig. 3.2. There are two electrons with energy E_1 and in the ground state there is one electron in state E_2. The dotted levels above E_2 represent permitted levels which the electron from E_2 may occupy if excited.

The energy-level system for a more complex multi-electron atom might have the general character shown in Fig. 3.3, where E_1, E_2, and E_3 are completely filled levels, and E_4 is the outermost shell or level which may be only partly filled.

The dotted levels are unoccupied levels into which electrons from E_4 may be excited. Apart from the innermost shell, E_1, the other levels

contain more than two electrons (e.g. eight) when they are completely filled. Energy levels E_2 and E_3 should thus each more properly be shown as several very closely spaced full lines, indicating that there is really a set of filled sub-levels.

Splitting of Energy Levels

If two identical atoms are brought close together, then they react upon each other and the energy-level picture is modified. In the new system, now containing two atoms coupled together, there is a tendency for each of the original permitted electron energy levels to be split into two. The amount of splitting depends upon the closeness of the atoms and the phenomenon is analogous to the coupling of two identical resonant electrical circuits to give a system with a double-humped frequency response the precise shape of which depends on the tightness of the coupling.

This type of coupling between atoms is responsible for the multiplicity of energy levels in molecules and the much greater complexity of their spectra compared with those of simple atoms. Transitions between closely spaced molecular energy levels give absorption and emission effects at microwave frequencies which have been discussed elsewhere.

In a solid crystal there are very many atoms close enough together to interact, and the single energy levels of the original atoms are broadened into wide bands of closely spaced levels.

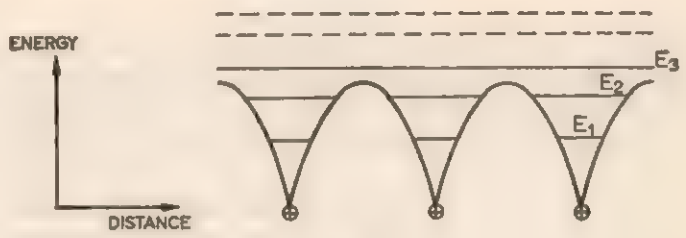

FIG. 3.4

Band System of Solids

If we consider a one-dimensional crystal composed of a single row of atoms, then the potential energy of an electron has the periodic form plotted as a heavy line in Fig. 3.4 due to the positively charged nuclei.

The horizontal lines shown in the figure are the electron energy levels in the isolated atom, full lines indicating occupied levels and dotted lines vacant levels.

This much simplified view of the energy levels in the solid suggests that those with energies E_1 and E_2 are confined to the immediate vicinity of the nuclei in the potential troughs, while those possessing energy E_3 can exist anywhere in the crystal. The electrons with energy E_3 are free to wander at random through the crystal lattice—they are the "free" electrons responsible for electrical conduction in early classical solid state theory.

A later, more comprehensive, theory takes into account the widening of the individual atomic levels into bands due to close packing. This theory explains most of the phenomena of current interest and we shall consider it below.

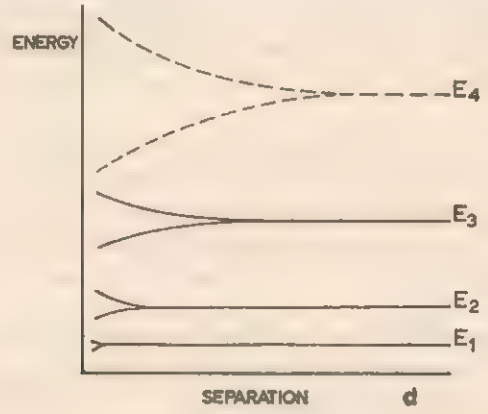

FIG. 3.5

Let us first consider two atoms separated by a distance d, each atom possessing the electron energy structure shown simplified as four levels, on the right-hand side of Fig. 3.5, for large values of d. We shall further assume that the element considered has levels E_1, E_2, and E_3 fully occupied.

E_1 thus contains two electrons of opposite spins, and E_4 is an unoccupied level to which an electron from E_3 could have gone if the atom had been excited. It has previously been mentioned that filled higher levels like E_2 and E_3 should be represented by a set of very close levels each containing two electrons of opposite spin. These sub-levels are, however, so close together that it is customary to show the energy of each shell as a single level.

When the two atoms are brought close together each individual energy level tends to split into two, the higher levels being affected first as the atoms approach and the amount of splitting increasing as the separation decreases (Fig. 3.5).

If the number of atoms reacting with each other is N and the average separation is d, then each of the single levels splits into N levels. An energy-level picture similar to Fig. 3.6 (i) would be obtained with the total separation of the N levels into which each original atomic level splits, increasing as the average atomic spacing decreases.

If the actual separation of the atoms in a particular crystalline solid is a (Fig. 3.6 (i)), then the electron energy level structure for that crystal will be as shown in Fig. 3.6 (ii). Each of the original levels of the isolated atom has been split into N levels, the higher bands now being of considerable width.

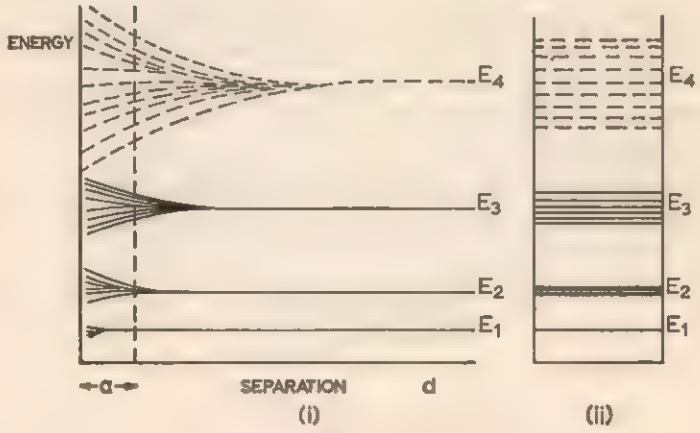

FIG. 3.6

In the isolated atom, E_1, E_2, and E_3 were completely filled levels, so the corresponding bands in the solid state are also completely filled. In particular $2N$ electrons completely fill the N levels of the band E_3 which is separated by a considerable energy gap from the band of vacant levels above it. These electrons are thus unable to accept a small amount of extra energy from a battery connected to the material because there are no permitted vacant levels near by. Since they cannot receive a small energy increment they cannot constitute a flow of current and the material is an insulator.

If in the original atom the level E_3 had been occupied by one instead of two electrons, then the N levels of the band E_3 in the solid would contain only N electrons. These electrons would tend to fill the band from the bottom upwards, so that E_3 would consist approximately of $N/2$ filled levels with $N/2$ vacant levels immediately above them.

The electrons in the lower half of E_3 would now be able to receive

a small increment of energy from an applied electric field and the material would be a conductor of electricity.

Overlapping Bands

The implication of the previous section is that an element will only be a conductor in the solid state if the highest electron energy level of the atom is incomplete, e.g. it contains one and not two electrons, so that the band evolved from it has only half the positions occupied. There is, however, a way in which the uppermost band of a crystal may be only partly filled even though the top energy level of the isolated atom is completely occupied.

In certain elements the levels split into wider bands than those indicated in Fig. 3.6 (i). At the atomic separation, which actually occurs in the crystal, all the vacant excited energy levels merge and this completely vacant band overlaps the completely filled band evolved from the highest fully occupied energy level.

The energy-level picture for the crystal will now have the form shown in Fig. 3.7 where the highest band, now partly filled, is due to an overlap

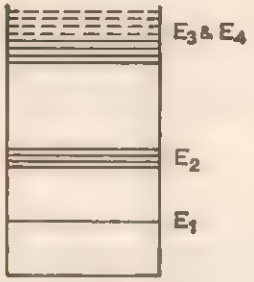

FIG. 3.7

between a filled and a vacant band. Note that where there are insufficient electrons to occupy the band completely, the levels are filled from the bottom up.

Conductors and Insulators

As long as we are only concerned with the electrical properties of the solid, then we need only consider the highest of the filled bands and the vacant band that lies above it. The energy-level pictures for all pure crystals then fall into one of the two classes shown in Fig. 3.8.

When there is no gap between the filled and vacant bands the material is a conductor, and when there is a forbidden energy gap between them

the material is an insulator because a small increment of energy is insufficient to take an electron from the filled valence band into a vacant level.

It is worth noting that energy-level diagrams, unless otherwise stated, are drawn for zero degrees absolute. At any real temperature an

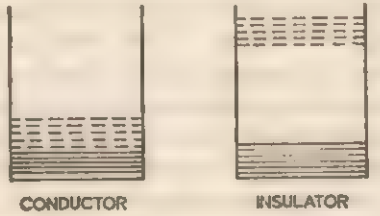

FIG. 3.8

electron may be thermally excited upwards into a slightly higher energy level if one is available. Thus some of the electrons near the top of the filled band in the conductor will be excited upwards into nearby vacant levels, leaving vacancies in the levels they previously occupied. Such thermal excitation will not occur in the insulator because of the width of the forbidden energy gap.

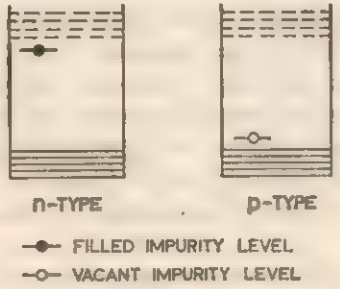

FIG. 3.9

Impurity Semiconductors

If a material in the pure form has the characteristic insulator structure, then the addition of small amounts of suitable impurity may change it into a semiconductor. Dependent on the kind of impurity added, the "doped" material may become either an *n*-type or a *p*-type semiconductor (Fig. 3.9).

In the *n*-type semiconductor the impurity gives rise to an extra

occupied electron energy level up near the conduction band. Such an effect is achieved by doping a tetra-valent element like silicon with a group five impurity like arsenic, or by preparing a group 3–group 5 compound like gallium phosphide so that it is rich in the group 5 element.

A classical view of the conduction process is that four of the valence electrons of the impurity atom are used to bind it into the tetra-valent crystal lattice, leaving the fifth as a "free" electron which allows conduction to take place. The energy level explanation is a little different.

The extra energy level due to the impurity atom will not be a general property of the crystal but will represent a situation existing only in the vicinity of an impurity atom. For this reason the impurity level is drawn as a short line to indicate that it is localised.

The situation in Fig. 3.9 is, however, that which occurs at zero degrees absolute, and at any higher temperature some of the impurity electrons will be excited up into the nearby conduction band. The filled or donor impurity level is usually so close to the conduction band that at room temperature all the impurity electrons will be excited into it. Once excited thermally into the conduction band the impurity electrons can exist anywhere in the crystal. They can also accept a small increment of energy, so the material will conduct electricity. The conductivity will depend on the number of electrons available in the conduction band, and since at room temperature all the impurity electrons are excited into this band, the conductivity will be proportional to on the number of impurity atoms present, i.e. the percentage doping.

A *p*-type semiconductor (Fig. 3.9) may be obtained if a tetra-valent material is doped with a trivalent impurity, e.g. indium, or if a group 3–group 5 compound is prepared in a form rich in the group 3 element. A vacant, or acceptor, impurity level or "hole" is produced down near the filled band and the energy level situation at zero degrees absolute is as shown.

At higher temperatures the hole is excited downwards into the valence band and is now free to exist anywhere in the crystal. It behaves as though it were charged positively and is responsible for the conduction in the *p*-type impurity semiconductor.

At 0° K each hole is localised in the vicinity of its impurity atom, but at higher temperatures one of the electrons from a neighbouring atom of the host material will occupy the hole, leaving a new hole by the host atom. This new hole will in turn be occupied by an electron from another atom, so that there appears to be a hole wandering at random through the material. The original impurity atom is now negatively charged but fixed in the crystal, while the hole wandering through the

crystal is effectively positively charged because its motion in one direction really represents the movement of an electron the same distance in the opposite direction.

The electrical situation at room temperature in a *p*-type semiconductor consists of a collection of stationary negatively charged impurity atoms distributed throughout the crystal and an equal number of positively charged holes wandering at random, so that the nett space charge is zero.

In the *n*-type semiconductor the impurity atoms become positively charged and are fixed in position, with electrons wandering at random, the nett space charge again being zero.

The behaviour of the *p*- and *n*-type semiconductor in contact is most important technologically. Devices such as rectifiers, transistors, tunnel diodes, and varactor diodes for parametric amplifiers all depend on the properties of the *p*–*n* junction.

Intrinsic Semiconductors

Certain solids possess an energy-level picture (Fig. 3.10 (i)) which has the characteristic insulator form, with the empty conduction band separated from the filled valence band by a forbidden energy gap. In this case, however, the energy gap is small and at temperatures above

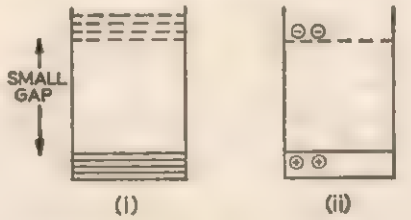

FIG. 3.10

0° K it is possible for an electron to be excited thermally across the gap into the conduction band, leaving a hole in the valence band. These intrinsic holes and electrons (Fig. 3.10 (ii)) can now act as charge carriers and the material will conduct electricity.

At any temperature hole–electron pairs will be produced at a certain rate in an intrinsic semiconductor and will also disappear at a certain rate when holes and electrons come together in the course of their random motion. At a given temperature there will be a certain number of charge carriers available, and hence a certain conductivity which increases with temperature.

Germanium, which is commonly used as a host material to which impurities are added to make impurity semiconductors, is also an intrinsic semiconductor. In such a germanium-based impurity semiconductor a significant amount of intrinsic current flows at room temperature. Devices like rectifiers and transistors depend upon impurity effects, and the presence of intrinsic current is a disadvantage, so that germanium transistors cannot be operated at temperatures above 70° C.

The unwanted effects of intrinsic semiconduction can be reduced by using a substance like silicon, with a larger energy gap, as the basic material.

The *p–n* Junction

It is possible to prepare a single crystal of a semiconductor so that one end of it is doped with *p*-type impurity atoms and the other end with *n*-type impurity atoms. The region in which the change from *p*- to *n*-type material occurs is called the junction and its character is different from that of the rest of the material. We shall consider the electrical character of a junction where the change from *p* to *n* takes place sharply but the same general properties arise in the graded junction where the transition is not so sharp.

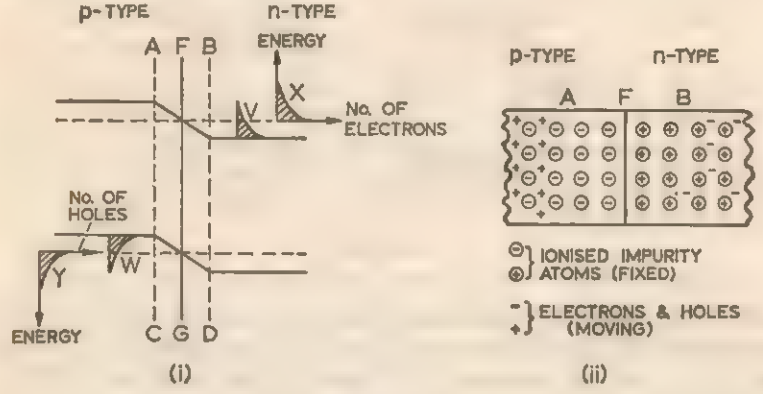

FIG. 3.11

Figure 3.11 (i) shows the energy-level picture at a sharp discontinuity *FG* between a *p*- and an *n*-type semiconductor. The horizontal dotted lines show the bottom of the vacant band and the top of the filled band in the positions they would have occupied if there had been no contact between the *n*- and the *p*-regions of the crystal.

At room temperature the impurity electrons in the *n*-type material and the holes in the *p*-type material will have been excited into the vacant band and the filled band respectively. Within these bands the electrons and holes will be distributed exponentially with energy as indicated by the shaded distribution curves at *X* and *Y*.

In Fig. 3.11 (ii) is shown the distribution of charge in the material: to the left of *A* and to the right of *B* is seen the charge distribution which would have existed in the *p*- and the *n*-type materials respectively if there had been no contact. In each case the charge due to the fixed ionised impurity atoms is just neutralised by the opposite charge of the holes or electrons which move at random through the material.

If we now imagine the contact to be made, then the electrons from the *n*-type material will diffuse into the *p*-type, and holes from the *p*-type will diffuse into the *n*-type. If this were only a diffusion process, then it would continue until the concentration of holes and electrons were constant throughout the crystal.

In fact, however, the electrons which leave the *n*-type material leave it positively charged, and the holes which leave the *p*-type material leave it negatively charged. They diffuse into the material in which they constitute minority charge carriers (e.g. electrons in *p*-type) until they meet a carrier of the opposite sign. When an electron and a hole meet they combine and both charge carriers are removed from the electrical conduction process. The immediate vicinity of the junction, say between *A* and *B*, contains no holes or electrons and is an insulator across which there exists a potential gradient due to the space charge of the ionised impurity atoms, now no longer neutralised by free carriers. The region between *A* and *B* is called the depletion layer.

The energy levels on each side of the junction move vertically with respect to each other because of the potential difference now existing between the *p*- and the *n*-type material. The final situation is shown by the full lines in Fig. 3.11 (i). Electrons have passed from *n* to *p* and holes in the opposite direction until a sufficiently great potential difference is established to prevent any further diffusion. When this has happened the hill in the energy level diagram is just high enough to prevent the most energetic electron in the *n*-type material—or hole in the *p*-type—from crossing the junction. The peak of the distribution curve *V* (or *W*) just reaches the top of the potential hill.

If an external voltage is applied to the junction, then the height of the hill is either raised or lowered and rectification occurs. If the external voltage is applied in the forward direction with the positive terminal to the *p*-type material, then the hill will be lowered and the tops of the distribution curves will project over it. Electrons will be injected from

n into *p* and holes from *p* into *n*, making a large forward current. With the external voltage applied the other way round, the hill is increased in size and no current flows, although in a real rectifier there is some reverse current due to intrinsic charge carriers.

Tunnel Diode

If enormously high concentrations of impurities are used in a *p–n* junction the potential hill formed without external bias voltage becomes so high that the bottom of the conduction band of the *n*-type material, with its impurity electrons, is depressed below the top of the valence band of the *p*-type, with its holes. The electrons are separated from permitted vacant states at the same level in the other material only by the depletion layer which is so thin that quantum mechanical tunnelling occurs through it.

As the forward bias is increased the conduction and valence bands of the two materials no longer overlap and the tunnel current is replaced by ordinary forward current. Such a diode is called a tunnel or Esaki diode and the *I–V* characteristic has a valley in it with a negative slope which may be used for amplification or oscillation. Such diodes are likely to be developed as low noise amplifiers at frequencies up to many Gc/s.

Gunn Effect and Other Properties

The greatest technological application of semiconduction has been the exploitation of *p–n* junction properties in the transistor, where a forward and a reverse biased junction are made to interact. The tunnel diode mentioned above represents a simplification, where a single junction may be made to give a negative resistance effect and thus used as an amplifier or oscillator. Similar oscillatory effects have been observed in other types of diode operated in special ways.

A great simplification in electronic devices might be achieved if such effects could be obtained in semiconductors without the need for special metallic contacts, or junctions between *p*- and *n*-type regions. In particular, at frequencies in the few Gc/s region, which is so important in modern communications, the mechanical tolerances on manufactured devices like klystrons and transistors are becoming very stringent, but at the same time the dimensions involved are becoming closer to those naturally available in crystal structures. If bulk amplifiers and oscillators can be found, then they are likely to replace elaborate manufactured devices because they will be easier and cheaper to make, more efficient at heat dissipation, faster in response time, and more robust.

The Gunn effect[1] is an important example of a bulk phenomenon which has already led to a practical device.

If an electric field is applied to a piece of gallium arsenide then a current flows. If this field is now increased the current also increases, as would be expected. But if the field is increased still further, then the mean current flowing decreases, thus showing a negative resistance effect which in principle might be used to give amplification or oscillation. There is, however, an extra feature to the current in the negative resistance region. In general, it shows very rapid fluctuations in value and in particular, if the length of the gallium arsenide across which the field is applied is made small, then the fluctuation in current occurs at a particular frequency in the Gc/s region.

The effects observed depend principally on the fact that the gallium arsenide has an energy level structure like that of Fig. 4.3 (p. 37). Electrons will normally be found in the lower valley of the conduction band, but they may transfer to the higher valley if given sufficient energy by, for example, a high electric field. It so happens that the effective mass of these "hot electrons" in the higher energy valley is greater than that of those in the lower valley, and hence the mobility of the hot electrons is less. If a high field causes electrons to move into the higher energy valley the current in the material will actually fall, because of the reduced mobility, and a negative resistance effect will occur.

When the field is high enough to produce hot electrons its distribution throughout the material no longer remains uniform. The fall in mobility due to the production of hot electrons at one point gives increased resistivity and hence a rise in the field in that region which will produce even more hot electrons and a further rise in the field. Since the total voltage available across the specimen is fixed, the field in the normal region of the gallium arsenide will fall, and so will the current.

The high field region is formed first at the cathode. It builds up rapidly and then moves across the specimen to the anode where it disappears. The field in the specimen then reverts to the uniform state for a short time until another instability builds up at the cathode and sets out to the anode.

If the specimen is long, disturbances set out from the cathode one after the other and the current fluctuations have a largely random character. If, however, the specimen is made short, only one disturbance at a time crosses from cathode to anode and the current fluctuations have a regular pattern at a frequency inversely proportional to the specimen thickness which is typically about 0·1 mm.

[1] "The Gunn Effect", Gunn, *International Science and Technology*, October 1965.

If one of the terminals is connected to the inner of a coaxial cavity then R.F. oscillations are produced in the cavity and both pulsed and c.w. operation have been achieved.[2] An experimental oscillator giving several milliwatts of c.w. power at 1 Gc/s has been reported[3] with prospects of improved performance in the near future.

[2] "Continuous Microwave Oscillations of Current in GaAs", Gunn *et al.*, *I.B.M. Journal*, November 1964.
[3] *Electronic Engineering*, September 1965.

4

The Interaction of Photons, Phonons, and Electrons

IN THE PREVIOUS CHAPTER we have examined what determines whether a substance is a conductor of electricity. The discussion was based on the energy-level structure of the solid and it was only necessary to consider the most energetic of the electrons present. The criterion for conductivity was found to be that these electrons should be able to accept small increments of energy from an applied electric field.

We have hardly mentioned what happens to the electrons once they are moving through the material under the influence of an electric field, although experience of the heating effect of an electric current suggests that they give energy to the bulk of the solid material.

It is thus possible for energy to be transferred from an electron to the crystal lattice. The converse process is also possible, i.e. energy may be transferred from the lattice to an electron. This converse process has already been met as the thermal excitation of electrons, e.g. across the energy gap in the intrinsic semiconductor (p. 25), or from a donor impurity level into the conduction band in the *n*-type impurity semiconductor.

Joule heating and thermal excitation both involve the interchange of energy between an electron in the material and the crystal lattice—both are "internal" energy processes.

There are also important "external" energy processes (e.g. in photoconductors, phosphors, and lasers) where an electron is excited by radiation entering the substance from outside, or, conversely, where an electron falls from a high to a lower energy level and emits radiation which escapes from the material.

The external energy interchanges are well known to be quantum processes involving bundles of energy $E = hf$, called quanta or photons. These interchanges are often known as photon–electron interactions.

Similarly, the internal energy interchanges are quantum processes.

The quantum of lattice energy involved is known as a phonon, by analogy with the photon, and the processes as phonon–electron interactions.

In this chapter we consider some of the properties of phonons, photons, and electrons, and the ways in which they interact.

Electrical Resistance

If an electric field, E, is applied to a free electron, then it experiences a force Ee and a constant acceleration Ee/m. The electron velocity thus increases continuously, as does the electron's energy, as long as it remains in the field.

This is certainly not the case when a current flows in a conductor due to an applied field. In this case the electrons appear to move with a constant drift velocity, and their mobility μ is defined as the average velocity acquired when a unit electric field is applied. There must therefore be some additional mechanism involved which extracts energy from the electrons as they move through the conductor. This extra mechanism is, in fact, the collision of the conduction electrons with the crystal lattice. The average time that the electron moves freely between collisions is sometimes called the relaxation time τ.

Mobility and relaxation time are related by the expression

$$\mu = \frac{e\tau}{m}$$

where m is more properly the effective mass of the electron which may differ from the true mass because of the special conditions under which the electron exists in the solid.

The conductivity of the material is given by $ne\mu$, where n is the number of conduction electrons per unit volume.

In a material where the conductivity is wholly or partially due to holes the same expressions may be used, but the values of μ and m will, in general, differ from those used for electrons.

Collisions with the Lattice

The simple view of the electron as a solid particle "colliding" with the equally solid ions of the crystal lattice gives a qualitative idea of the origin and nature of electrical resistance. It is, however, unable to account for all conductivity phenomena, especially the variation with temperature. In particular superconduction, the complete disappearance of resistance in certain materials at very low temperatures, seems to imply the disappearance of the crystal lattice, which is clearly absurd.

To get a full explanation of the conduction process the electron must

be regarded as a wave packet interacting with the periodic potential field of the lattice. Such a wave-mechanical treatment can be found in many advanced textbooks on the solid state.[1]

In the wave-mechanical treatment a "collision" becomes the random scattering of the electron wave and it is shown that a perfect periodic potential structure does not affect the propagation. It is only when the periodicity of the potential becomes disturbed that scattering occurs and electrical resistance appears.

Such disturbances of the perfect periodicity of the lattice occur as a result of thermal vibration at any temperature above the absolute zero. These vibrations are, in fact, elastic waves propagated in the solid at the velocity of sound, which is much less than the velocity of the electron waves. Consequently, the electron wave effectively passes through a stationary lattice with its regularity instantaneously distorted by the elastic wave.

Disturbances of the periodicity of the lattice, and hence scattering, may be due not only to thermal vibration but to distortion caused by foreign atoms occurring either accidentally in impure materials, or by design in alloys.

Material of the highest purity must be prepared if the best conductivity is to be obtained, e.g. 0·1% of phosphorus impurity in copper reduces the conductivity by about 50%. Physical distortion in the lattice can be caused by strain due to working the material, and the conductivity may be improved several per cent by annealing. If the material is prepared as an alloy for reasons of mechanical strength, then a fall in conductivity must usually be accepted, e.g. a copper–cadmium alloy containing a little less than 1% cadmium has a conductivity just less than 90% of that of pure copper.

Phonons

There are a number of phenomena in physics which can be described either in terms of waves or particles, although in any particular problem one approach may be more suitable than the other.

The controversy about the wave and corpuscular nature of light was a lengthy historical argument and the explanations of interference effects in terms of wave properties, and of the photo-electric effect in terms of photons, are both well known. The photo-electric effect is concerned with the amount of energy $E = hf$ in one photon, but by using the Einstein $E = mc^2$ relation it is possible to ascribe to the photon an effective mass and momentum which must be used to explain the Compton scattering of X-rays (p. 42).

[1] *Wave Mechanics of Crystalline Solids*, R. A. Smith, Chapman & Hall.

In the case of the electron its particle-like properties of mass and momentum were long known and used before its dual character was also established. Its wave properties are the basis of the wave mechanics which is a powerful tool in the investigation of the solid state, and one of the results of which we have already used in noting above that a perfectly periodic crystal lattice offers no resistance to electric current.

One of the principal sources of electrical resistance is the distortion of the lattice due to thermal vibrations, and this provides us with a third example of duality.

The distortions travel through the material as elastic waves and this vibrational energy is quantised, just as light energy travelling as an electromagnetic wave is quantised. By analogy with the photon we can consider the passage of vibrational energy through the solid as due to phonons moving through the material at the velocity of sound. The phonon is thus a quantum of lattice energy and, like the photon, it will possess momentum and mass.

Electrical resistance can now be reconsidered as due to the collision of electrons with phonons (and with impurities). As the temperature rises, increased atomic vibrations can be represented by increasing numbers of phonons flowing through the crystal. More collisions with phonons will now occur, giving reduced relaxation time and mobility and hence an increase in resistance with temperature.

We shall see later that phonon momentum may be important in certain electron–photon interactions in solids.

Thermal Conductivity

If one part of a substance is maintained at a high temperature and another at a low temperature, then energy flows from the hot to the cold region. It may be convenient to think of the energy being transported by particles, and these "particles" will be phonons in non-metallic solids, and phonons and electrons in metals.

The thermal conductivity is proportional to the mean free path of the particles, and the mean free path of electrons is much greater than that of phonons. Consequently, the thermal conductivity of non-metals is very low compared with that of metals, although the presence of foreign atoms, either as impurity or in alloys, sharply reduces metallic thermal conductivity.

There is obviously great similarity between the mechanisms of thermal and electrical conductivity, and in the case of metals this is summed up in the Wiedemann–Franz law which suggests, with fair accuracy, that the ratio of the thermal to the electrical conductivity is the same for all metals at a given temperature.

Electron–Photon Interactions

The phenomena of electrical and thermal conductivity discussed above are largely due to electrons and phonons, and the collisions between them. In solid state electronics there is often emission of quanta of radiation when electrons undergo energy changes. Such electron–photon interactions in solid semiconductors concern us particularly because of their occurrence in the crystal injection laser, and are of general interest in the emission of light by phosphors. The converse processes involving absorption of quanta of radiation are perhaps of less interest in practical electronics, although the study of these processes yields valuable information about the solid state.

The simplest possible approach to the emission of a photon is that which deals only in conservation of energy.

In Fig. 4.1 is shown the energy-level structure for a semiconductor

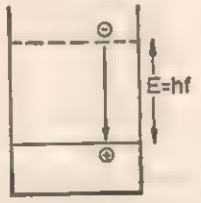

FIG. 4.1

with a hole and an electron, produced by some sort of earlier excitation, in the valence and conduction bands respectively. The hole and the electron will be wandering at random through the material. If they meet then they will recombine, disappear as charge carriers, and a photon of radiation will be emitted.

Assuming the electron to fall from the bottom of the conduction band into a hole at the top of the valence band, then the frequency of the radiation emitted is given by $E = hf$, where E is the energy gap in the material. This energy gap must be large if visible light is to be emitted.

Such a simple process is called direct recombination, but in many cases the recombination is complicated by extra levels due to imperfections or impurities.

In Fig. 4.2 (i) an impurity level, in this case a vacant level, is shown near the edge of the conduction band.

The recombination is now a two-stage process illustrated in Fig. 4.2 (ii) and (iii).

If the electron arrives at one of the sites in the crystal where there is

an impurity level, then it falls into that level and remains fixed in position in the material (ii). Subsequently, one of the holes moving through the material will arrive at this site and be trapped by the temporarily filled impurity level (iii). The level is now in its original state and again available to assist in the recombination of a free hole and electron.

If both processes, i.e. the original capture of an electron and the subsequent capture of a hole, are carried out with equal probability by the impurity level then it is called a recombination centre. A similar recombination centre may be due to the existence of a filled instead of a vacant impurity level, but in this case the first process will be the capture of a hole and the second the capture of an electron.

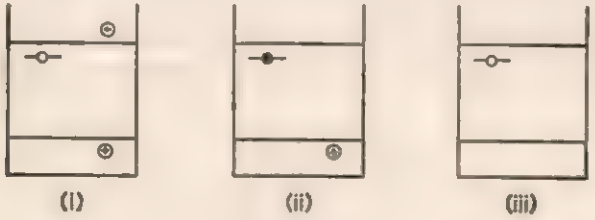

FIG. 4.2

If any impurity centre does not perform each of the two processes of recombination with equal probability, then it may be known as a trap. Thus a vacant impurity level may readily capture an electron, but, when filled, have a much lower probability of completing the recombination by capturing a hole. Such an impurity level would be known as an electron trap.

Usually the free holes and electrons produced in a material by some form of excitation will recombine fairly quickly, emitting the excess energy at the appropriate frequency. But if trapping centres exist in the material, say electron traps, then some of the excited electrons will be trapped and held for a time in a state of comparatively high energy. They will eventually recombine with free holes and emit energy, but this may not occur until some time after the original excitation is removed. The material will thus behave as a phosphor with "after-glow" properties.

Features of the material which determine the way in which energy is emitted will be the total width of the energy gap, and whether the impurity level is near the conduction or valence band (shallow), or approximately in the middle of the gap (deep).

A complete description of the possible recombination processes and an explanation of the character of the energy emitted would obviously

be very complex. Indeed, the details of the processes involved are not yet fully understood. One important consideration must, however, be added to the simple conservation of energy principle used above.

The principle that momentum as well as energy must be preserved in crystal processes is fundamental theoretically, and has important implications in applied quantum electronics.

In discussing the conservation of crystal momentum in electron–photon processes in semiconductors it is useful to consider the energy–momentum diagram for a crystal. A very simple theoretical example in one dimension is shown in Fig. 4.3.

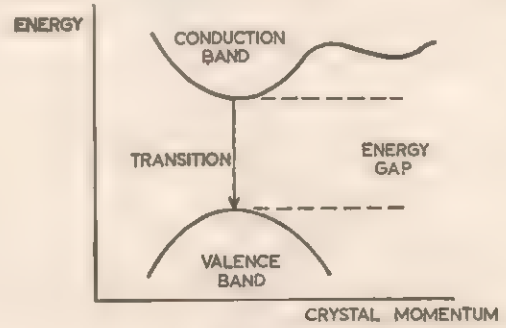

FIG. 4.3

Here the upper line represents the conduction band and the lower line the valence band, and there is a forbidden energy gap between them. Electrons are most likely to be found at the conduction band minimum and holes at the valence band maximum. Since the momentum of a photon in or near the visible spectrum is so small there must be negligible change in crystal momentum when an electron falls from the conduction band to the valence band and emits a photon of energy.

In the diagram shown, such a transition is readily possible and the material will be known as a direct gap semiconductor (e.g. gallium arsenide).

More often the energy–momentum diagram will have the form shown in Fig. 4.4, where in general the transition of an electron across the gap will involve a considerable change in crystal momentum. The electron can only make the direct transition shown, and radiate a photon of energy, if the extra momentum is removed simultaneously as a phonon. This process is not very probable and in such materials the light-emitting transition of an electron across the gap will be indirect and will make use of an impurity level in the gap. The impurity centre must be such that it absorbs a phonon of the right momentum, but it

must not be so deep in the energy gap that the subsequent transition does not produce visible light, i.e. if the material is required as a crystal lamp or injection laser (p. 129), then luminescent impurity centres of this type must be encouraged, and other, non-radiative centres avoided.

Indirect gap materials are silicon and germanium, but their energy gaps are so small that light is not emitted. A luminescent indirect gap material, such as gallium phosphide, will emit light of a colour which depends upon the type of added impurity.

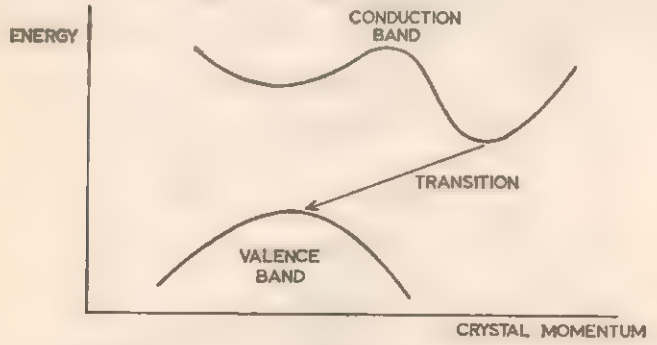

FIG. 4.4

A radiative transfer of an electron from the conduction to the valence band of a direct gap semiconductor is an electron–photon interaction, whereas the same process in an indirect gap material usually proceeds by way of an impurity centre and can be considered as an electron–phonon–photon interaction.

In electrical conduction the phonon is a crude spoiler, limiting by collisions the velocity acquired by the electrons, but in luminescence it plays a useful role by allowing radiative transfers to occur in indirect gap materials. We shall see elsewhere (p. 59) another example of the phonon being useful: in the phonon maser.

Absorption—Excitons

The previous section dealing with electron–photon interactions assumed that excitation of electrons into the conduction band had already taken place. It then considered how radiative transitions of these electrons back across the gap to the valence band would occur. The converse process of absorption of a photon of radiation to produce an electron in the conduction band and a hole in the valence band must obey the same rules of conservation of energy and momentum. As with luminescence, the way in which the electron transition takes place will be

greatly influenced by the nature of the energy–momentum diagram of the material and may require the emission or absorption of a phonon.

Apart from the fact that radiation absorption in solids follows the same rules as emission it is of much less practical interest to the electronics engineer. Much of the fundamental knowledge of the solid state, however, comes from absorption experiments carried out in the physics laboratory.

The width of the energy gap in a semiconductor can be determined by observing the maximum light wavelength which will cause conductivity. Such experiments are best carried out at low temperatures where there will be few phonons to collide with electrons in the valence band and excite them into the conduction band where they will produce unwanted intrinsic conduction effects which might mask the photoconductivity which is to be observed.

When such experiments were carried out with light of gradually increased wavelength it was found that for a given material there was a certain wavelength at which the absorption suddenly fell. This maximum wavelength showed the width of the energy gap. But at certain still longer wavelengths there was a significant amount of absorption indicating the presence of extra levels in the gap. Such levels could be due to impurities, and indeed this absorption technique can be used to determine the position of the impurity levels. If impurity levels are the cause of the discrete absorption at longer wavelengths, then it will be accompanied by photoconductivity due to the mobile charges produced in either the conduction or the valence band.

There is, however, another phenomenon which produces selective absorption at certain wavelengths longer than the expected maximum value. This is known as the exciton process.

In the simple case of the excitation of an electron from the valence to the conduction band a hole is produced in the valence band. Ordinarily the hole and the electron move off in different directions through the material as quite separate carriers of charge and thus contribute to the electrical conductivity. In an analogous way, a hydrogen atom given enough energy would be ionised and the electron and the positive nucleus would go their different ways.

If, however, the hydrogen atom had been given less than the energy required to produce ionisation, then it might have been merely raised to one of the discrete set of excited energy levels. In a similar fashion a solid may absorb a photon which is of insufficient energy to excite an electron across the gap and thus produce a separate hole and electron. The photon may, however, create a hole–electron pair still bound together, but in an excited state like the hydrogen atom. This bound

hole–electron pair may move through the material and is called an exciton.

The exciton gives rise to extra energy levels which lie within the energy gap of the material. They are not, however, like vacant impurity levels because, although the electron may be excited into one of them, it is still bound to the hole and no conductivity results.

Excitons may thus contribute to the absorption or emission of energy in a solid but not to the conductivity.

More Complex Interactions

There is still much to be learned about the solid state and the interactions between electrons, photons, phonons,[2] and excitons discussed briefly above are some of the simplest which are needed to explain important general principles. The detailed explanation of many phenomena involves more complex interactions, perhaps the emission of radiation by an exciton, or the emission of many phonons by an excited trap.

Perhaps the most important of these more complex interactions are the so-called non-linear photon processes which are now being discovered and investigated using the very high energy fluxes available from lasers. Typical of these processes is the excitation of an electron from the valence to the conduction band by the simultaneous absorption of two or more photons, not necessarily of the same frequency. Another non-linear process is the simultaneous absorption of two or more identical photons and the emission of a single photon at a harmonic of the original photon frequency. Thus the techniques of frequency doubling or harmonic generation common in conventional tuned circuit electronics may be applied at the highest radio frequencies and in the optical region.

Accounts of such new phenomena and applications of recently discovered ones appear regularly in the scientific journals. The periodicals quoted in the bibliography (p. 141) are typical of those which carry such reports.

[2] "Phonons and Phonon Interactions", *B.A.K.*, published Benjamin (1964).

5

Radiation and Matter—Einstein's Radiation Theory

IN THE EARLIER DISCUSSION of the interactions between photons and electrons in solids it was found that the principle of conservation of energy was not enough and that conservation of momentum had also to be taken into account (p. 37). Nevertheless the momentum of the photon in these interactions was usually very small compared with that of the electron, and it was the momentum of a phonon which was relied upon to balance the momentum equation if required.

In this chapter, however, we are concerned principally with the relations between radiation and matter, and we shall find that the momentum associated with a photon of electro-magnetic radiation must be taken into account if certain important processes are to be explained. Some such processes of fundamental and historical importance are briefly dealt with below, but one of especial interest is considered more fully. This is Einstein's investigation of fundamental radiation properties which led to the recognition of the phenomenon of stimulated emission and, eventually, to the use of this process in masers and lasers.

Compton Effect

The early years of the twentieth century saw the reopening of the old discussion about the dual nature of electromagnetic radiation. This time, however, the wave and corpuscular theories were not in conflict, but it was recognised that the choice of either to explain a given phenomenon was largely a matter of convenience. By 1924 the duality concept had been extended to matter by de Broglie, who suggested the association of a wavelength $\lambda = \frac{h}{mv}$ with a moving particle. Conversely, at about the same time (1923), Compton explained a particular form of X-ray scattering by associating momentum with electromagnetic radiation.

In certain cases the X-rays scattered from a material have a longer wavelength than the incident radiation and the change in wavelength $d\lambda$ depends on the scattering angle ϕ.

$$d\lambda = \frac{h}{mc}(1 - \cos\phi)$$

Compton scattering (Fig. 5.1) is due to a photon of X-radiation striking an electron. The electron recoils with kinetic energy $\frac{1}{2}mv^2$ at an angle θ to the path of the incident photon, and the scattered photon moves away at an angle ϕ to the incident path with a diminished energy. The scattered radiation is of longer wavelength than the incident radiation, since $hf = hf' + \frac{1}{2}mv^2$, i.e. energy is conserved.

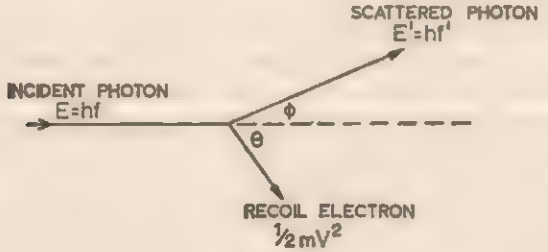

FIG. 5.1

It is, however, possible to observe experimentally both the direction of the scattered X-radiation and that of the recoil electron. To explain the relation between θ and ϕ, and to derive the relation between $d\lambda$ and ϕ, it is necessary to consider conservation of momentum as well as conservation of energy.

Because of the equivalence $E = mc^2$ between mass and energy, a photon, moving with the velocity of light c, has an effective momentum hf/c. Two equations for the conservation of momentum may thus be derived

$$\frac{hf}{c} = \frac{hf'}{c}\cos\phi + mv\cos\theta$$

$$\frac{hf'}{c}\sin\phi = mv\sin\theta$$

Combined with the energy equation $hf = hf' + \frac{1}{2}mv^2$ these equations enable $d\lambda = \lambda' - \lambda$ to be determined in terms of ϕ, i.e.

$$d\lambda = \frac{h}{mc}(1 - \cos\phi)$$

Since $d\lambda$ is independent of wavelength, the fractional change in wavelength, which will determine how easily the effect may be observed, becomes greater as the wavelength decreases. In general, Compton scattering is only observed with the high energy, high momentum photons in the X-ray and γ-ray region.

Mössbauer Effect

The Compton effect was an early demonstration of the momentum of radiation and was consequently very important theoretically, although of negligible practical importance. The Mössbauer effect is a very much more recent (1959) demonstration of the momentum of radiation and is thus intrinsically not of such great theoretical importance as the Compton effect. It is, however, of great practical importance because it provides a tool for experimental work of considerable interest in fundamental scientific theory.

The nucleus of a radioactive element which emits γ-rays only will have two energy states. A transition from the high state to the low state will involve the emission of a photon of γ-radiation. If E_2 and E_1 are the energy levels, then the frequency of the γ-radiation is given by $E_2 - E_1 = hf$. The width of the spectral line emitted (cf. p. 15) will increase with the probability of such emission occurring, and a narrow γ-ray line will hence be emitted by a radioactive material of relatively long half-life.

In the case of γ-radiation it is more common to define the photon energy than its frequency, and for a particular element Fe_{57} the energy of the emitted γ-radiation is about 14,000 eV and the line width about 10^{-8} eV.

Suppose that the nucleus emitting the radiation is free to recoil, then in order that momentum be conserved the nucleus will move backwards with momentum equal to that of the photon. There will be some energy associated with the nucleus as it recoils, and this much less energy will be available for radiation. The energy emitted will thus be hf' instead of hf and the frequency of the radiation $f' < f$.

In the same way a nucleus could normally absorb energy hf and become excited from state E_1 to state E_2. But, if the nucleus is free to recoil on absorbing the photon of radiation, then the incident photon will have to possess rather more energy than hf in order to provide the recoil energy and the excitation energy $E_2 - E_1$. The frequency of the radiation absorbed by the nucleus will thus be f'' where $f'' > f$.

In a material where the nuclei are free to recoil then the γ-ray emission frequency differs from the absorption frequency by $f'' - f'$.

The energy associated with the recoil of the iron nucleus is about

10^{-3} eV and the line width of the radiation only 10^{-8} eV, so that the absorption and emission spectra do not overlap.

If, however, the nuclei are set in a solid crystal lattice, usually cooled, then there is negligible energy lost in recoil and the absorption and emission frequencies coincide. In such a case the absorption by unexcited nuclei of radiation emitted by excited ones can readily be detected. Because of the extremely narrow line width anything which causes the emission frequency to change slightly will result in a fall in absorption.

Such a change in frequency can be caused by the Doppler effect, and relative velocities between source and absorber of just a few feet per hour can be detected.

A change in frequency is predicted by the general theory of relativity if a photon moves in a gravitational field. Only the Mössbauer experiment is sensitive enough to detect this effect. If a source is placed at the top of a tower and the absorber at the bottom, then the change in absorption observed is of the correct order. More accurate experiments along these lines in the near future are expected to provide stringent tests for the relativity theory.

Einstein Theory of Mechanism of Radiation

The Maxwell curve for the distribution of velocities among gas molecules and the Planck curve for the spectral distribution of thermal radiation are very similar in shape. Wien had made use of this similarity in deriving his pre-quantum expression

$$E_f = f^3\phi\left(\frac{E}{T}\right) \quad . \quad . \quad . \quad . \quad \text{(A)}$$

where E_f is the energy density at temperature T of thermal radiation of frequency f.

Einstein tried to use Wien's original arguments plus the fundamental assumptions of the quantum theory in order to derive the Planck thermal radiation expression. In the course of his derivation[1] Einstein found it necessary to make certain assumptions about the fundamental processes involved in the absorption and emission of radiation. Because the derivation was successful it must be assumed that these hypotheses were correct.

Einstein's derivation is followed below in order to draw attention to these fundamental radiation processes and to establish certain quantitative relations between them. Both the fundamental laws and the

[1] "On the Quantum Theory of Radiation", Einstein, *Phys. Zeits.*, Vol. 18, 1917, pp. 121–8.

quantitative relations are of great importance in the theory and design of masers and lasers.

The discussion is essentially concerned with the energy associated with molecules, and two relations based on the quantum theory must at once be stated.

First, only certain discrete states 1, 2, 3, . . . n, are permitted by the quantum theory and these will have energies E_1, E_2, E_3 . . . E_n.

Second, the relative frequency of occurrence (W_n) of the state n at temperature T is given by the statistical relation

$$W_n = p_n\, e^{E_n/kT} \quad . \quad . \quad . \quad . \quad \text{(B)}$$

(see also p. 3)

where k is Boltzmann's constant and p_n, the weight of the state, is independent of temperature and is characteristic of the nth state.

To these two relations Einstein now added his three basic hypotheses about the radiation process. We shall consider these three hypotheses in relation to two states of the molecule characterised by energy E_1 and E_2, where E_2 is greater than E_1, the molecule being able to pass from one state to the other by the emission or absorption of radiation at frequency f.

(1) The probability that in time dt a molecule in the high energy state E_2 will drop to the lower energy state E_1 by the emission of energy $E_2 - E_1$ as radiation of frequency f is given by

$$dW = A_{21}\, dt \quad . \quad . \quad . \quad . \quad \text{(C)}$$

where A_{21} is a characteristic constant for the two levels considered and can be called the spontaneous emission coefficient.

(2) In the presence of radiation of frequency f and density E_f the probability that a molecule in state E_1 will in time dt absorb energy and rise to state E_2 is given by

$$dW = B_{12}E_f\, dt \quad . \quad . \quad . \quad . \quad \text{(D)}$$

where B_{12} is a constant and can be called the absorption coefficient, or sometimes the induced absorption coefficient.

(3) The converse process to (2). In the presence of radiation of frequency f and density E_f the probability that a molecule in state E_2 will in time dt be induced to emit energy and fall to state E_1 is given by

$$dW = B_{21}E_f\, dt \quad . \quad . \quad . \quad . \quad \text{(E)}$$

where B_{21} is a constant and can be called the induced emission coefficient.

These are Einstein's three elementary radiation processes and they are now applied to a collection of molecules at a temperature T.

From the statistical expression (B) concerning the frequency of occurrence of a given state

$$W_n = p_n \varepsilon^{-E_n/kT}$$

we can write down the relative frequency of occurrence of the two states with energy E_1 and E_2 as

$$p_1 \varepsilon^{-E_1/kT} \quad \text{and} \quad p_2 \varepsilon^{-E_2/kT}$$

Now using (C), (D), and (E) above we can obtain the number of each of the three radiation processes occurring in time dt.

Number of spontaneous emission processes is

$$p_2 \varepsilon^{-E_2/kT} \,.\, A_{21}\, dt$$

Number of induced emission processes in the presence of radiation of density E_f is

$$p_2 \varepsilon^{-E_2/kT} \,.\, E_f \,.\, B_{21}\, dt$$

Number of absorption processes in the presence of radiation of density E_f is

$$p_1 \varepsilon^{-E_1/kT} \,.\, E_f \,.\, B_{12}\, dt$$

If these radiation processes are not to destroy the equilibrium state then the number of emission and absorption processes must be equal.

$$p_2 A_{21} \varepsilon^{-E_2/kT}\, dt + p_2 B_{21} E_f \varepsilon^{-E_2/kT}\, dt = p_1 B_{12} E_f \varepsilon^{-E_1/kT}\, dt$$

thus
$$p_2 \varepsilon^{-E_2/kT}[A_{21} + B_{21}E_f] = p_1 \varepsilon^{-E/kT} B_{12} E_f \,. \quad . \quad \text{(F)}$$

But the energy density of thermal radiation depends upon the temperature and if T becomes infinite then so does E_f,

i.e.
$$p_2[A_{21} + B_{21}E_f] = p_1 B_{12} E_f$$

and
$$p_2 B_{21} = p_1 B_{12} \quad . \quad . \quad . \quad . \quad \text{(G)}$$

Substituting this relation back into (F) gives

$$p_2 \varepsilon^{-E_2/kT}[A_{21} + B_{21}E_f] = p_2 B_{21} \varepsilon^{-E_2/kT} E_f$$

$$E_f p_2 B_{21}[\varepsilon^{-E_1/kT} - \varepsilon^{-E_2/kT}] = p_2 \varepsilon^{-E_2/kT} A_{21}$$

$$E_f \frac{B_{21}}{A_{21}} = \frac{\varepsilon^{-E_2/kT}}{\varepsilon^{-E_1/kT} - \varepsilon^{-E_2/kT}}$$

$$E_f = \frac{A_{21}/B_{21}}{\varepsilon^{E_2 - E/kT} - 1}$$

This expression becomes identical with Planck's radiation distribution expression (p. 4)

$$E_f = \frac{8\pi h f^3}{c^3}\left(\frac{1}{\varepsilon^{hf/kT} - 1}\right)$$

if $E_2 - E_1$ equals hf

and
$$\frac{A_{21}}{B_{21}} = \frac{8\pi h}{c^3} f^3$$

$\frac{8\pi h}{c^3} f^3$ is sometimes written as $\frac{8\pi f^2}{c^3} hf$ or $N_M hf$, where N_M is the number of modes per unit volume (p. 3).

Einstein thus succeeded in his attempt at an alternative derivation of Planck's expression and the hypotheses he adopted about the nature of the interaction of radiation and matter, (1), (2), (3), must therefore receive strong support.

In developing his argument this far Einstein had considered only the energy involved in radiation processes, but now he extended it to include the momentum transfer taking place (cf. p. 37).

> ". . . in general, we are content to limit ourselves to the consideration of the *energy* exchange, without taking into account the exchange of *momentum*. We can easily feel justified in so doing, since the smallness of the momentum transferred by the radiation makes this momentum insignificant in actuality in comparison with other causes of motion. But for the *theoretical* treatment, even this small effect must be considered just as important as the immediately obvious transfer of *energy* by radiation, because energy and momentum are so closely tied up with one another. Hence a theory can be regarded as valid only after it has been shown that the momentum which, according to this theory, is transferred from radiation to matter leads to the type of motion which the theory of heat demands."

Einstein made the following assumptions about the transfer of momentum to a molecule when it is irradiated or when it spontaneously emits a quantum of energy.

If a pencil of radiation acts so that a molecule encountered by it takes in or gives out a quantum of energy hf, then momentum hf/c will be transferred to the molecule—in the direction of propagation of the pencil if energy is absorbed by the molecule, and in the opposite direction if energy is emitted by the molecule.

If a molecule gives out a quantum of energy by the spontaneous emission process, then it will be emitted in a random direction and the

corresponding impulse suffered by the molecule will be oppositely directed.

Using these assumptions he then showed that the motion thus imparted to the molecules would lead to exactly that distribution of velocities which was given by the kinetic theory of gases.

Powerful additional support was thus given to the original hypotheses, (1), (2), (3), and also to these additional ideas about momentum interchange. We shall now examine the significance of these ideas in maser and laser theory.

If we are interested in amplification, then we must consider principally the effects (2) and (3) which occur when radiation is incident upon the material. If induced emission occurs, then the beam of radiation is enhanced, and if induced absorption occurs, then the beam is attenuated. Thus, for amplification we require as much stimulated emission and as little absorption as possible.

Suppose that in the material there are N_2 units in the higher energy state E_2, and N_1 units in the lower energy state E_1.

Then the number of induced emission processes in unit time is

$$N_2 B_{21} E_f$$

where B_{21} is the induced emission coefficient and E_f is the density of the incident radiation.

The number of absorption processes in unit time is

$$N_1 B_{12} E_f$$

where B_{12} is the induced absorption coefficient.

Each of these processes involves the emission or absorption of the same amount of energy $E_2 - E_1 = hf$, so that for amplification to occur

$$N_2 B_{21} E_f > N_1 B_{12} E_f$$

i.e.

$$N_2 B_{21} > N_1 B_{12}$$

But from (G) above

$$p_2 B_{21} = p_1 B_{12}$$

and if we consider the simplest possible case where the difference in the statistical weighting factors p_1 and p_2 can be neglected, then $B_{21} = B_{12}$.

The condition for amplification thus becomes $N_2 > N_1$ and the material must be maintained in a non-equilibrium state with more units in the higher than in the lower energy state. The normal equilibrium state is when $N_2 = N_1 e^{-(E_2 - E_1)/kT}$, so that energy must be supplied to the material in such a way that units in state E_1 are pumped up to state E_2. The non-equilibrium state when $N_2 > N_1$ is sometimes also

referred to as population inversion, or as negative temperature distribution, and the achievement of such states is discussed in the next chapter.

In the electronic processing of any information an important consideration is the signal to noise ratio. The noise which may be generated in an amplifier, and superimposed on the signal while it is passing through, is of especial importance if weak signals are to be handled.

Spontaneous emission is a random process unrelated to the incident field and thus represents noise generated in the amplifier.

It has already been shown that the ratio of the spontaneous emission coefficient to the stimulated emission coefficient is given by

$$\frac{A_{21}}{B_{21}} = \frac{8\pi h f^3}{c^3}$$

Thus the noise generated in a stimulated emission amplifier will rise rapidly relative to the amplification produced as the signal frequency rises. Although such an amplifier used in the microwave region (maser) has an exceptionally good noise performance, the same would certainly not be true of a similar amplifier used at optical frequencies (laser). In fact, the laser is not likely to be generally used as an amplifier, certainly not of weak signals, but rather in the oscillator form as a generator of coherent light.

6

Non-Equilibrium Systems

IF THE INDIVIDUAL ELEMENTS of any system can possess different amounts of energy, then there will normally be more with low energy than there are with high energy. A non-equilibrium system is one in which this normal state of affairs has been inverted, so that the population in a high energy level is greater than that in a lower energy level.

Einstein has shown (p. 48) that in a two-level system the probability of a quantum of radiation being absorbed by a unit in the low energy level is equal to the probability of the radiation causing the emission of an extra quantum of similar radiation from a unit in the high energy level. If there are more units in the high level than in the low one, i.e. if inversion has been achieved, then there will be more emission than absorption and amplification will occur.

In this chapter we shall examine the Einstein concept of stimulated emission, and the behaviour of the non-equilibrium system, in order to see how the two ideas may be brought together to achieve maser and laser action.

Stimulated Emission

The basis of maser and laser action is the stimulated emission of radiation, which was first recognised by Einstein in 1917[1] as one of the processes involved in the interaction of matter and radiation.

He found a derivation of Planck's black body radiation expression which made use of the similarity in form between Planck's radiation distribution curve and Maxwell's curve of velocity distribution. The derivation was obtained by considering that the distribution of energy required by the quantum theory could only arise through the absorption and emission of radiation.

Certain hypotheses about the absorption and emission of radiation by molecules formulated in the course of the derivation greatly clarified

[1] "Zur Quantentheorie der Strahlung", Einstein, *Phys. Zeit.*, Vol. 18, 1917 (also p. 44).

ideas about the interaction of radiation and matter. These ideas are conveniently considered in terms of a system of units with two permitted energy levels $E_2 > E_1$, such that transitions occur between them with the emission or absorption of a quantum of energy $E_2 - E_1 = hf_{21}$.

Three processes are possible in such a system:

(i) A unit in the high energy level E_2 may at any time emit radiation hf_{21} and fall to the state E_1. This process will occur whether or not there is a beam of radiation of frequency f_{21} passing through the material. The quantum of radiation may be emitted in any direction and with any phase. The process is thus random and is called spontaneous emission.

The second and third processes can only occur when a beam of radiation at frequency f_{21} is passing through the material.

(ii) A unit in the low energy state E_1 may absorb a quantum of radiation at frequency f_{21} from the beam and rise to state E_2. In passing through the material the original beam of radiation will be attenuated by this absorption process. Algebraically, one may consider that photons of the same phase and direction as those in the original beam are subtracted from it.

(iii) A unit in the high energy state E_2 may be induced by an incident quantum to emit another identical quantum of radiation at frequency f_{21}, in the same direction and with the same phase as the original quantum. The original beam of radiation will thus be amplified by this induced emission process. Algebraically, photons of the same phase and direction as those in the original beam are added to the beam and the induced emission process is the converse of the absorption process.

There are two important quantitative results relating to these three processes which are used below.

The first is that the absorption and induced emission probabilities are equal. $W_{12} = W_{21}$.

The second is that the probabilities of spontaneous and induced emission are in the ratio $(e^{hf_{12}/kT} - 1)$ so that in the optical spectrum at room temperatures the ratio is very large, and spontaneous emission predominates, whereas in the microwave region stimulated emission predominates.

Population Inversion—Non-equilibrium State

If spontaneous emission is ignored, then the fact that the absorption and induced emission probabilities are the same ($W_{12} = W_{21}$) means that amplification occurs when the population N_2 of the higher energy state E_2 is greater than the population N_1 of the lower state E_1.

Such a state of population inversion was achieved in the first maser[2] by making use of the two energy states of the ammonia molecule which are separated by an energy gap corresponding to a frequency of 24 Gc/s.[3] An ammonia molecule in the higher energy state is repelled, and one in the lower energy state attracted, by a strong electrostatic field. So by passing a stream of molecules through a suitable electrostatic separator it is possible to produce a gas with $N_2 \gg N_1$.

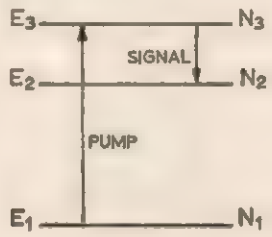

FIG. 6.1

Physical separation of the high and the low energy units in this fashion is, however, not common and the principles of population inversion in the maser are best brought out by considering a three-level system operating in the microwave region.[4]

The choice of the microwave region allows two simplifications to be made in the theory. First, spontaneous emission may be neglected, and second, the linear approximation to the Boltzmann expression $e^{-hf/kT} = 1 - \frac{hf}{kT}$ may be used, since at 3 Gc/s $\frac{hf}{kT} \simeq 4 \times 10^{-4}$.

Signal amplification is to be obtained at frequency $f_{32} = \frac{E_3 - E_2}{h}$, and in order to do so a population inversion $N_3 > N_2$ is required. This inversion is achieved by supplying separately a large amount of energy at frequency f_{31}, the pump frequency, so that large numbers of units from state E_1 are raised to state E_3.

Under normal thermal equilibrium conditions before pump power is applied, $N_3 = N_1 e^{-hf_{31}/kT} = N_1\left(1 - \frac{hf_{31}}{kT}\right)$.

Thus when the pump is first applied there will be more units which absorb energy than are induced to emit it. Upward transitions between

[2] "The Maser", Gordon *et al.*, *Phys. Rev.*, Vol. 95, 1954.

[3] "E-m Waves of 1·1 cm Wavelength and the Absorption Spectrum of Ammonia", Cleeton and Williams, *Phys. Rev.*, Vol. 45, 1934.

[4] "Proposal for New Type Solid State Maser", Bloembergen, *Phys. Rev.*, Vol. 104, 1956.

E_1 and E_3 will exceed downward transitions until eventually N_3 would become equal to N_1.

This assumes that transitions between the states are due only to the absorption or emission induced by the pump power. In the solid state, however, transitions between levels may be induced by interaction between the active units and the thermal energy (phonons) of the host lattice in which they are embedded.[5] Since the active units in the maser are usually paramagnetic atoms with the levels, E_1, E_2, E_3, due to electron spin, this process is called spin-lattice interaction or relaxation.

With practical maser materials the probability of spin-lattice interactions can be kept low (spin-lattice relaxation time high) so that near saturation pumping is, in fact, achieved and N_1 and N_3 tend to become equal when pump power is applied.

If the spin-lattice transition probabilities are denoted by w_{13}, w_{31}, w_{12}, etc., then we can derive certain relations between them for use later by considering the state of thermal equilibrium when lattice induced transitions between states will be equal, i.e. $w_{32}N_3 = w_{23}N_2$, $w_{21}N_2 = w_{12}N_1$, etc.

But in thermal equilibrium

$$N_2 = N_1\left(1 - \frac{hf_{21}}{kT}\right) \quad \text{and} \quad N_2 = N_3\left(1 + \frac{hf_{23}}{kT}\right)$$

using the Boltzmann approximation.

$$\text{Thus } w_{12} = w_{21}\left(1 - \frac{hf_{12}}{kT}\right) \quad \text{and} \quad w_{32} = w_{23}\left(1 + \frac{hf_{23}}{kT}\right)$$

When signal (f_{23}) and pump (f_{13}) power are applied, then, in addition to the above lattice induced transitions, radiation induced transitions occur. Let W_{23} W_{32} W_{13} W_{31} be the probability of such transitions.

Then because induced absorption equals induced emission

$$W_{23} = W_{32} \quad \text{and} \quad W_{13} = W_{31}$$

Let us first consider a general case where radiation induced and lattice induced transitions may occur between all levels.

$$\frac{dN_2}{dt} = N_3(W_{32} + w_{32}) + N_1(W_{12} + w_{12}) - N_2(W_{21} + W_{23} + w_{21} + w_{23})$$

In the particular case we are considering, radiation is available only at frequency f_{32} (signal) and f_{31} (pump), so that W_{21} and W_{12} do not appear.

Also from above $W_{23} = W_{32}$ and $W_{13} = W_{31}$.

[5] "Masers", Weber, *Rev. Mod. Phys.*, Vol. 31, No. 3, 1959.

$$\text{Thus } \frac{dN_2}{dt} = W_{32}(N_3 - N_2) + (N_1 w_{12} - N_2 w_{21}) + (N_3 w_{32} - N_2 w_{23}).$$

$$\text{But from above } w_{12} = w_{21}\left(1 - \frac{hf_{12}}{kT}\right) \text{ and } w_{23} = w_{32}\left(1 - \frac{hf_{32}}{kT}\right)$$

$$\therefore \frac{dN_2}{dt} = W_{32}(N_3 - N_2) + w_{32}(N_3 - N_2) + w_{21}(N_1 - N_2) + \frac{N_2 h f_{32} w_{32}}{kT} - \frac{N_1 h f_{12} w_{21}}{kT}$$

In the steady state $\frac{dN_2}{dt} = 0$, so approximately

$$N_3 - N_2 = \frac{Nh}{3kT}\left(\frac{w_{21}f_{21} - w_{32}f_{32}}{W_{32} + w_{32} + w_{21}}\right)$$

where $$N_1 + N_2 + N_3 = N$$

So that for population inversion between states E_3 and E_2, and consequently signal amplification at frequency f_{32}, $w_{21}f_{12} - w_{32}f_{32}$ must be positive. The lower the temperature the more favourable the non-equilibrium state.

If, as is often the case, the energy-level structure is roughly symmetrical, so that $f_{12} \doteqdot f_{32}$, then there must be considerable disparity in w_{21} and w_{32}, the spin-lattice transition probabilities, if a reasonable degree of inversion is to be achieved.

The signal power emitted by the maser will be $(N_3 - N_2)W_{32}hf_{32}$ and it can be shown that W_{32} increases if the interaction between the active units is small, i.e. if the spin–spin transition probability $\left(\frac{1}{\tau_2}\right)$ is small, i.e. if τ_2 the spin–spin relaxation time is high.

The power output can, however, be seen to depend upon the concentration of active units N. But as N increases so τ_2 decreases, so that there will be an optimum concentration for maximum power output.

In the first solid state maser, paramagnetic gadolinium ions were chosen as the active units because a convenient steady magnetic field gave an energy-level structure with transitions in the microwave region. The host lattice, chosen for its magnetic inertness, was lanthanum ethyl sulphate, and the concentration of gadolinium was 0·5%.

The pump frequency was 17·5 Gc/s and the signal frequency 9 Gc/s, so that there was little disparity between f_{21} and f_{32}. For good inversion it was therefore necessary that w_{21} and w_{32} should differ by as much as possible. It was found that doping the crystal with 0·2% cerium atoms[6] gave a ten to one disparity between the two spin-lattice transition probabilities and allowed maser action to be observed.

[6] "Use of Active Material in Three-Level Solid State Masers", Schulz-DuBois *et al.*, *Bell Syst. T.J.*, Vol. 38, 1959.

Choice of Active Medium

Having seen what is meant by population inversion and the influence upon it of transition probabilities, it is worth examining some other general principles which must be considered in choosing an active material for a maser or laser. To a certain extent the choice of material will be influenced by the convenience with which it can be pumped, but this aspect of the choice is deferred until the next section.

The first consideration in choosing an active material is that it shall have an energy-level structure which permits transitions emitting energy of the frequency at which amplification or oscillation is required.

Transitions at microwave frequencies can occur in free atoms and molecules, and some of these have been discussed in Chapter 2. Some of the gases in which such transitions take place occur naturally, like oxygen and water vapour in the atmosphere and like atomic hydrogen in interstellar space. Other gases like caesium and ammonia have been used in microwave devices like the atomic clock and the first maser. Conventional spectroscopy shows that there is a multitude of transitions in the optical range available in atoms in the gas and vapour states. The main problem in obtaining laser action is not the finding of a suitable energy-level system but the devising of a method of pumping. The noble gases, mercury, and water vapour are some of the materials in which laser action has been obtained.[7]

The energy-level schemes of atoms closely packed in the solid state becomes exceedingly complex because of interaction between the separate atomic systems (Chapter 3). In general, simple solid materials do not give line spectral transitions, although the presence of impurities may give absorption and emission at discrete frequencies. In particular, ions of the rare earth elements suitably dispersed as impurities in crystals, glasses, or even as organic compounds in solution, will allow the establishment of population inversion and show laser action.[8] If a suitable steady magnetic field is applied to certain single crystals doped with these rare earth ions, then microwave transitions and maser action can be produced, with the added advantage of tuning by varying the applied magnetic field.

In any active medium there are conflicting requirements about the concentration of the units. A high concentration is required if the gain

[7] "Gaseous Optical Masers", Bennett, *Applied Optics, Supplement 1*, 1962.
[8] "Solid State Lasers", Davies and Moore, *I.E.E. Laser Symp.*, September 1964.
[9] "Masers", Weber, *Rev. Mod. Phys.*, Vol. 31, No. 3, July 1959.

is to be high, while too high a concentration gives unwanted interaction between the units. In the gas devices, usually lasers, the concentration of active units is small so that the gain per unit length is also small and a very long resonant cavity, in fact, a Fabry–Perot system about a metre long, must be employed. Solid lasers and masers, very frequently ruby, have a higher concentration of active elements and with a higher gain can thus occupy a much smaller volume in either a cavity or a travelling wave system. It is, however, much more difficult to prepare the solid maser material as it must usually be cut from a perfect single crystal. Although the solid laser in transmitting systems can give very high pulse power, heating effects spoil the laser action and the mean power output is often no better than that of the gas laser, which is normally used if continuous operation is required, although CW operation of ruby lasers has been achieved.[10]

Pumping Methods

In general for amplification to occur in a maser-type system an inversion of population must be achieved between the two levels which give a transition at the frequency to be amplified. Most practical microwave systems achieve this inversion by using a three-level system of the general type proposed by Bloembergen and discussed on p. 52.

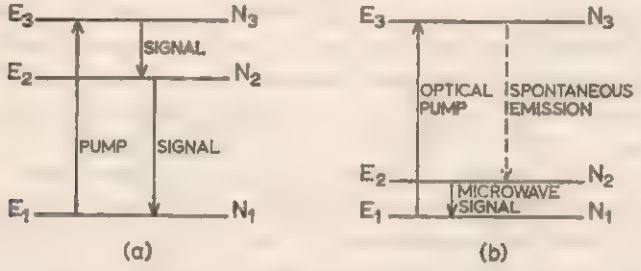

FIG. 6.2

Such a three-level system is shown in Fig. 6.2 (*a*) and, depending on the material chosen, it could operate either with signal due to transitions between levels 3 and 2, in which case population inversion between 3 and 2 would be required, or with signal due to transition between levels 2 and 1, in which case population inversion between 2 and 1 is required.

In Fig. 6.2. (*b*) a similar three-level system is shown, but the gap between E_2 and E_3 is now very large and pumping from E_1 to E_3 is due

[10] "A continuously Operating Ruby Optical Maser", Nelson and Boyle, *Applied Optics, Supplement 1*, 1962.

to optical energy supplied to the material. The microwave transition and the population inversion are between levels 2 and 1.

This optical pumping of microwave masers shows promise of improving and extending their performance.[11] It enables the low noise performance of microwave amplifiers like the ruby maser to be realised at higher temperatures (e.g. liquid nitrogen) thus avoiding the necessity for very elaborate refrigeration. If a microwave pumped maser is operated at such a temperature the gain-bandwidth product and the noise performance are both poor, whereas optical pumping preserves the gain-bandwidth and noise performance and will also allow the maser to be operated at much higher frequencies.

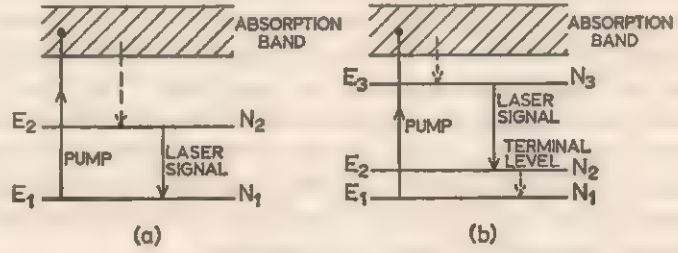

FIG. 6.3

In the case of lasers there are more pumping methods available than there are for microwave masers. We shall first discuss the optical pumping method which takes place in a fairly standard three- or four-level system in much the same way as has already been described.

This optical pumping method is normally used in solid lasers. In gas lasers other more elaborate gas discharge pumping methods are available involving collision between electrons and the gas, or between different atoms or molecules in gaseous mixtures; these methods are described later.

The energy-level schemes used in the optical pumping of solid lasers are shown in Fig. 6.3. In each case the pumping is from the ground state E_1 to a broad absorption band which is characteristic of the laser material, followed by a radiationless, phonon assisted transition from this broad band to a nearby sharp level E_2 in (*a*) and E_3 in (*b*). The laser transition then takes place between this sharp level and E_1 the ground state in (*a*) or E_2 the terminal level in (*b*). In each case a wide-band source of light such as a xenon flash tube or a mercury discharge is used as the pump.

[11] "Optical Pumping of Microwave Masers", Hsu and Tittel, *P.I.E.E.*, Vol. 51, January 1963.

In the three-level system (*a*), of which ruby is an example, a population increase is established between levels 2 and 1. A better system (*b*) establishes the inversion between level 3 and a terminal level 2 which is significantly higher than the ground state from which the pump is working. The pump thus has the large population of level 1 to pump from but only has to build level 3 up to a greater population than that of the terminal level 2. The population in level 2 will be very much less than that in 1 if the terminal level is a fair distance above the ground state. The establishment of inversion will occur for very much lower pump power, especially if the system is cooled. Calcium fluoride doped with uranium and cooled in liquid nitrogen uses a four-level system.[12]

In pumping a gas laser the fact that the uppermost energy level is not a broad band as in Fig. 6.3 but sharp like the other levels, means that only a very small fraction of the energy of a broad band pumping source is efficiently used. Optical pumping of a gas laser is only used if the frequency of the pumping transition required in the active material happens to coincide with the frequency of a strong line in the emission spectrum of some other gas. Under these circumstances quite efficient pumping can be achieved and such a system is used in the caesium laser pumped by light emitted from a helium discharge.

The active element in a gas laser system is nearly always pumped into the high energy level by one of the many different sorts of collision process which may occur in a gas discharge.[13] This type of process can provide reasonably efficient pumping because a fairly wide band of incident particle energies can produce the required transition in the target atom or molecule, the excess energy being carried away as kinetic energy.

In the simplest case electrons in the discharge excite the gas atoms to the required high energy level; such a process with consequent population inversion may occur in a discharge in a pure noble gas. A second possible process involves a mixture of gases such as helium and neon. The population inversion occurs in the neon, but the original excitation is due to collision between electrons and helium and the subsequent transfer of the energy the helium has gained to the neon. There is a further set of processes where a molecule is given enough energy in a collision to dissociate into a pair of atoms, one of which is in the required excited state.

The establishment of population inversion in a gas discharge is a very complex process with the possibility of several different contributing reactions in a particular case. Much research work is concentrated on

[12] *Lasers*, Lengyel, Wiley, 1962.
[13] "Gaseous Optical Masers", Bennett, *Applied Optics, Supplement 1*, 1962.

understanding these processes and investigating new gases and mixtures in which inversion might be achieved.

Other Considerations

The general principles of non-equilibrium systems discussed in this chapter grew from Einstein's radiation studies (Chapter 5) and are themselves extended into the applied field in Appendices I and II. A quite different method of achieving a non-equilibrium system is discussed in connection with the injection laser (Appendix V), and the phonon maser,[14] which is mainly of theoretical interest, uses an ultrasonic wave to produce inversion.

It is perhaps worth noticing that in this chapter it has been established that inversion can lead to amplification, but little has been said about the usefulness of such amplification. Applications are dealt with in the appendices, but it is appropriate in concluding this chapter to point out that spontaneous emission has not been discussed. The fact that spontaneous emission, and hence noise, increases so rapidly with frequency is responsible for the practical devices being split roughly into microwave amplifiers and optical oscillators. Oscillation, of course, implies the existence of amplification, but whereas the maser is a simple microwave amplifier invaluable for its low noise properties, the corresponding light amplifier would be of little use because of its very high noise figure. In the optical region the amplification principle is put to use indirectly to give oscillation and hence provide lasers, the high power coherent sources of light previously unobtainable.

[14] "The Phonon Maser", Tucker, *Proceedings*, Paris 1963 Quantum Electronics Conference, Columbia University Press, 1964.

7

Power Relations in Parametric and Non-Equilibrium Systems

IN THE MASER AMPLIFIER we have already described a system where a signal at one frequency may be amplified if energy at another frequency is supplied to a suitable substance with the appropriate passive electrical circuits. Such amplifiers depend upon the upsetting of the normal equilibrium distribution of molecules among the various permitted energy levels of the active substance and are called non-equilibrium systems.

In addition to the maser there is another class of amplifier which is supplied with energy at one frequency in order to amplify a signal at some other frequency. This other type of amplification is called parametric and, like the maser, is important because of its low noise figure. The active element of a parametric amplifier is typically, but not always, a condenser whose capacitance can be made to vary with time.

Although the parametric amplifier has been discussed theoretically for over a century and is essentially easier to make than the maser, it was the maser which was first produced as a working amplifier.

Having already discussed the maser in some detail, we shall in this section state the most general properties of the parametric amplifier and derive certain important general principles which are common to both types. A discussion of practical parametric devices will be undertaken later.

Types of Parametric Amplifier

In the parametric amplifier, as in the maser, there are two frequencies introduced into the system from external sources. These are the signal frequency and the frequency of the pump which supplies power to the system in order that amplification may be achieved.

The action of the time-varying parameter of the amplifier is to provide frequency mixing so that in the general case there will be generated the

sum and the difference of the signal and the pump frequencies, and the sum and difference of all the harmonics of the signal and pump frequencies. The frequencies generated may be written as

$$\sum_{m=0}^{\infty} \sum_{n=-\infty}^{\infty} (m\omega_1 + n\omega_2)$$

where m and n are integers, and ω_1 and ω_2 are the original frequencies.

In the practical parametric amplifiers of most interest the electric circuits are arranged so that only three of the multiplicity of possible frequencies can exist in the system. These three frequencies are ω_1, ω_2 and $\omega_1 + \omega_2$ or $\omega_1 - \omega_2$. Amplifiers which have power flowing at $\omega_1 + \omega_2$ are called upper side-band amplifiers and those which have power flowing at $\omega_1 - \omega_2$ are called lower side-band amplifiers.

Of the three frequencies present in the amplifier, one is that at which the signal enters (ω_s) and is normally fixed by factors outside the control of the designer of the amplifier. The second frequency, that of the pump (ω_p), can be chosen by the designer quite freely, but once this is chosen then the third frequency can only have one of the two values $\omega_p \pm \omega_s$. A suitable circuit to allow power to flow at the chosen frequency $\omega_p + \omega_s$ or $\omega_p - \omega_s$ must be incorporated in the amplifier.

The frequency at which the low-power signal enters the amplifier is ω_s and in many amplifiers this is the frequency at which the high-power amplified signal leaves the amplifier. Provided that power amplification occurs, there is often no objection to the signal leaving the amplifier at a frequency different from ω_s. In the case of the three-frequency amplifier considered above, four different arrangements appear to be possible:

(i) Third or idler frequency $\omega_p - \omega_s$. Signal out at ω_s.

(ii) Third or idler frequency $\omega_p - \omega_s$. Signal out at $\omega_p - \omega_s$.

(iii) Third or idler frequency $\omega_p + \omega_s$. Signal out at ω_s.

(iv) Third or idler frequency $\omega_p + \omega_s$. Signal out at $\omega_p + \omega_s$.

In fact, only the arrangements (i), (ii), and (iv) give amplification and are used in practice.

We require to find a set of general principles which will show us why arrangement (iii) is useless, and to give us some idea of the sort of performance we can expect from the useful arrangements (i), (ii), and (iv). The Manley–Rowe relations are a set of principles which provide this information and further guidance on the feasibility or otherwise of projected parametric systems of greater complexity, e.g. one allowing more than three frequencies to flow.

Manley–Rowe Relations

The Manley–Rowe relations[1] were originally derived for a system containing a non-linear reactance and a number of circuits tuned to particular frequencies. The results, however, are equally applicable to the maser, and Weiss[2] has pointed out that the derivation of the relations is very much simpler in the quantum mechanical case.

In order to see simply and quickly the form and usefulness of the Manley–Rowe relations we shall first consider the particular three-frequency system described above and shall use a quantum mechanical treatment to obtain the power absorbed or generated at the various frequencies.

Power Relations in the Three-frequency System

We shall consider an amplifier in which only three frequencies can exist. In a parametric amplifier this would mean that energy could only flow at three frequencies determined by the three tuned circuits contained in the system. Similarly, in the corresponding maser, energy could only flow at the three frequencies determined by the three permitted energy levels of the maser material.

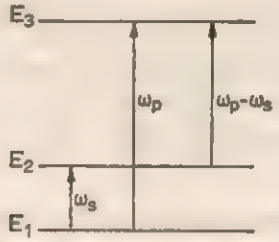

FIG. 7.1

Let two of the three frequencies in the system be ω_s and ω_p, the frequencies at which signal and pump power enter the amplifier. The third or idler frequency must now be chosen as $\omega_p - \omega_s$ or $\omega_p + \omega_s$.

We shall first choose $\omega_p - \omega_s$ when the three-frequency system in maser terms becomes as Fig. 7.1.

E_1, E_2, and E_3 are the three energy levels, and under steady conditions the population of each level will remain constant.

[1] "Some General Properties of Non-Linear Elements", Manley and Rowe, *P.I.R.E.*, Vol. 44, July 1956.

[2] "Quantum Derivation of Energy Relations Analogous to Those for Non-Linear Reactances", Weiss, *P.I.R.E.*, Vol. 45, July 1957.

There will be energy absorbed and liberated at all three frequencies, and each time a quantum of energy is absorbed or liberated one unit of the population will leave one level and go to another.

Let N_{12} denote the *nett* number of units per second going from E_1 to E_2 with N_{13} and N_{23} referring to the nett number of transitions per second from E_1 to E_3 and E_2 to E_3 respectively.

If the population in E_2 is to remain constant

$$N_{12} = N_{23} \quad . \quad . \quad . \quad . \quad (1)$$

But each transition from E_1 to E_2 means absorption of one quantum of energy $E_2 - E_1 = hf_s = \frac{h}{2\pi}\omega_s$.

Similarly, each transition from E_2 to E_3 means absorption of energy $\frac{h}{2\pi}(\omega_p - \omega_s)$.

Thus total power absorbed at signal frequency is

$$P_s = N_{12}\frac{h}{2\pi}\omega_s$$

Total power absorbed at idler frequency $\omega_p - \omega_s$ is

$$P_i = N_{23}\frac{h}{2\pi}(\omega_p - \omega_s)$$

Thus $\quad N_{12} = \frac{2\pi}{h} \cdot \frac{P_s}{\omega_s}$ and $N_{23} = \frac{2\pi}{h} \cdot \frac{P_i}{\omega_p - \omega_s}$

But from (i) above $N_{12} = N_{23}$ and thus

$$\frac{P_s}{\omega_s} = \frac{P_i}{\omega_p - \omega_s} \quad . \quad . \quad . \quad . \quad (A)$$

Similarly, if the population of E_3 is to remain constant, then

$$N_{13} + N_{23} = 0 \quad . \quad . \quad . \quad . \quad (2)$$

If we now consider the energy per quantum in transitions between E_1 and E_3 and betwen E_2 and E_3 we shall obtain a relation involving the pump power.

Power absorbed at pump frequency is

$$P_p = N_{13}\frac{h}{2\pi}\omega_p$$

Thus $N_{13} = \frac{2\pi P_p}{h\ \omega_p}$ and we have already seen that $N_{23} = \frac{2\pi}{h}\frac{P_i}{\omega_p - \omega_s}$.

But from (2) above $N_{13} + N_{23} = 0$, and thus

$$\frac{P_p}{\omega_p} + \frac{P_i}{\omega_p - \omega_s} = 0 \quad . \quad . \quad . \quad \text{(B)}$$

But since $N_{12} = N_{23}$, a third relation can be written

$$\frac{P_p}{\omega_p} + \frac{P_s}{\omega_s} = 0 \quad . \quad . \quad . \quad . \quad \text{(C)}$$

Let us now collect together the three relations (A), (B), and (C) and examine some of their significance in the design and operation of practical parametric systems.

It is well to remember that we are dealing with a particular type of amplifier in which the third frequency has been chosen to be $(\omega_p - \omega_s)$. This frequency will be referred to as ω_i, the idler frequency, as a convenient shorthand.

$$\frac{P_s}{\omega_s} = \frac{P_i}{\omega_i} \quad . \quad . \quad . \quad . \quad \text{(A)}$$

$$\frac{P_p}{\omega_p} + \frac{P_i}{\omega_i} = 0 \quad . \quad . \quad . \quad . \quad \text{(B)}$$

$$\frac{P_p}{\omega_p} + \frac{P_s}{\omega_s} = 0 \quad . \quad . \quad . \quad . \quad \text{(C)}$$

In our case energy is absorbed from the pump, which is the external power supply, so P_p is positive. Thus from (C) P_s must be negative, which means that more power is generated at the signal frequency than is absorbed from the signal source, i.e. there is power gain.

Because the idler frequency is chosen to be $\omega_i = \omega_p - \omega_s$ this is called a lower side-band amplifier, and because the amplified output is at the same frequency as the signal (ω_s) it is called a straight amplifier.

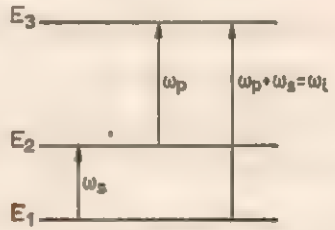

FIG. 7.2

Suppose that we had decided to choose an idler frequency of $\omega_p + \omega_s$, this is then called an upper side-band amplifier and it can be represented as in Fig. 7.2.

Using the same nomenclature as in (1) and (2) above

$$N_{12} = N_{23} \quad \text{and} \quad N_{13} + N_{23} = 0$$

$$P_s = N_{12} \frac{h}{2\pi} \omega_s$$

$$P_p = N_{23} \frac{h}{2\pi} \omega_p$$

Hence

$$\frac{P_s}{\omega_s} = \frac{P_p}{\omega_p} \quad . \quad . \quad . \quad . \quad . \quad \text{(D)}$$

$$P_i = N_{13} \frac{h}{2\pi}(\omega_p + \omega_s) = N_{13} \frac{h}{2\pi} \omega_i$$

Thus

$$\frac{P_i}{\omega_i} + \frac{P_p}{\omega_p} = 0 \quad . \quad . \quad . \quad . \quad \text{(E)}$$

Since

$$N_{12} = N_{23},\ N_{13} + N_{12} = 0$$

and

$$\frac{P_i}{\omega_i} + \frac{P_s}{\omega_s} = 0 \quad . \quad . \quad . \quad . \quad \text{(F)}$$

Let us now collect the three relations (D), (E), (F), and see how this upper side-band amplifier can be made to provide gain.

$$\frac{P_s}{\omega_s} = \frac{P_p}{\omega_p} \quad . \quad . \quad . \quad . \quad \text{(D)}$$

$$\frac{P_i}{\omega_i} + \frac{P_p}{\omega_p} = 0 \quad . \quad . \quad . \quad . \quad \text{(E)}$$

$$\frac{P_i}{\omega_i} + \frac{P_s}{\omega_s} = 0 \quad . \quad . \quad . \quad . \quad \text{(F)}$$

In all amplifiers P_p is positive because the pump supplies power to the amplifier. From (D) P_s must also be positive, which means that the amplifier absorbs more energy from the signal source than it liberates at the signal frequency. There is thus no possibility of power gain at the signal frequency and an upper side-band straight amplifier is not feasible.

Power is absorbed at the pump and signal frequencies, but from (F)

$$-P_i = P_s \frac{\omega_i}{\omega_s}$$

so that power is generated at the idler frequency. Since $\omega_i = \omega_p + \omega_s$ P_i is greater in magnitude than P_s and amplification of signal power is

possible with the output now at a different and higher frequency than the input signal. This amplifier is accordingly called an upper side-band up-converter.

The power gain $\left(G = \frac{P_i}{P_s}\right)$ is $\frac{\omega_i}{\omega_s} = \frac{\omega_p + \omega_s}{\omega_s} = 1 + \frac{\omega_p}{\omega_s}$, and for a given ω_s a large value of pump frequency is chosen if large gain is required.

Since $P_p = P_s \frac{\omega_p}{\omega_s}$, a large pump frequency implies that a high pump power will also be needed. This is to be expected because high power gain means that the whole system is operating at a high-power level and will thus draw high power from its main energy source, the pump.

We have seen that the upper side-band amplifier cannot be a straight amplifier but can give power gain as an up-converter where the power output is taken at the idler frequency. The lower side-band amplifier has already been shown to be capable of straight amplification and now examination of relations (A), (B), (C) will show that it can also give power gain with the output at the idler frequency.

$$\frac{P_p}{\omega_p} + \frac{P_i}{\omega_i} = 0$$

Thus $-P_i = \frac{\omega_i}{\omega_p} P_p$ and power is developed at the idler frequency. Since $\omega_i = \omega_p - \omega_s$, ω_i may be greater or less than ω_s and the amplifier may be an up- or a down-converter.

The three possible arrangements to obtain power amplification in a three-frequency parametric amplifier may be summarised thus:

Upper side-band amplifier

Idler frequency $\omega_i = \omega_p + \omega_s$.

Output at ω_i, hence up-converter.

$-P_i = \frac{\omega_i}{\omega_s} P_s$. Power gain $\frac{\omega_s}{\omega_i}$, which is always finite, so amplifier is stable.

Lower sideband amplifier (i)

Idler frequency $\omega_i = \omega_p - \omega_s$.

Output at ω_s, hence straight amplifier.

$-P_s = \frac{\omega_s}{\omega_p} P_p$. Power output independent of power input at signal frequency, hence gain indeterminate. Unstable and may oscillate at signal frequency.

Lower side-band amplifier (ii)

Idler frequency $\omega_i = \omega_p - \omega_s$.

Output taken at ω_i. If $\omega_i > \omega_s$, amplifier is up-converter; if $\omega_i < \omega_s$, it is down-converter.

$-P_i = \frac{\omega_i}{\omega_p} P_p$. Power output independent of power input at signal frequency hence gain is indeterminate. Unstable and may oscillate at idler frequency.

Practical Significance of Power Relations

Let us consider in simplified equivalent circuit form the properties summarised above for the three amplifiers.

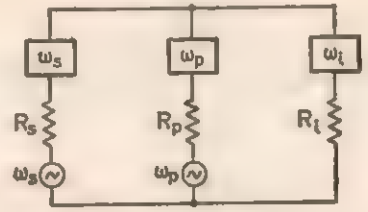

FIG. 7.3

In Fig. 7.3 the filters shown as square boxes limit the power flow in the three resistors R_s, R_p, and R_i to the appropriate frequencies ω_s, ω_p, and ω_i. There are two generators in the system, the signal source outside the amplifier and the pump inside the amplifier.

The resistor R_s, in fact, represents the internal resistance r_o of the signal source, the loss resistance r_s of the signal circuit, and, in addition, R_L an effective load resistor if power is taken out at the signal frequency.

If power is extracted from the idler circuit, then the load resistor appears as part of R_i, which now consists of the loss resistance r_i of the idler circuit plus R_L.

Straight amplifier:	$R_s = r_o + r_s + R_L$	$R_i = r_i$
Up- or down-converter:	$R_s = r_o + r_s$	$R_i = r_i + R_L$

The power relations summarised for the three amplifiers in the previous section hold for ideal conditions, e.g. lossless circuits. The inter-relation of r_o, r_s, R_L, etc., are important in determining the performance of a practical amplifier.

For example, the lower side-band amplifiers give output power, either at signal or at idler frequency, which is independent of the input signal power and there is thus a possibility of instability and oscillation

at the output frequency which may be ω_s or ω_i depending on the type of amplifier.

The rest of the amplifier connected to the output circuit can be regarded as a negative resistance because it is feeding power into the circuit. The magnitude of the negative resistance in the lower side-band amplifiers is controlled by the pump power. As long as the total resistance in the circuit has a nett positive value, then there is stable gain. If, however, in attempting to increase the gain, the total resistance is allowed to become zero or negative, then oscillation occurs.

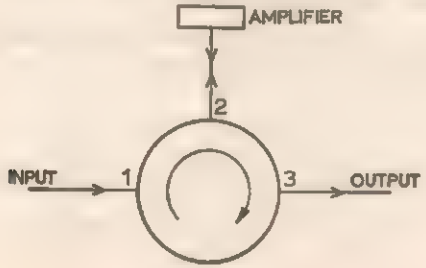

FIG. 7.4

In the straight amplifier the input signal and the amplified output both appear in the same circuit, so that the amplifier is a "one-port" device, i.e. the input and the output both use the same terminal. The difficulties of circuit organisation which arise with a one-port amplifier can be overcome conveniently by the use of a circulator, Fig. 7.4, at the sort of frequencies met in low noise applications (p. 127).

Degeneracy

In the lower side-band amplifier $\omega_i = \omega_p - \omega_s$, and it may be an up- or a down-converter depending on whether ω_i is greater or less than ω_s.

If ω_p is chosen to be equal to approximately $2\omega_s$, then ω_i and ω_s will be different but nearly equal.

In particular it can be arranged that the signal and idler frequencies are close enough together for both of them to lie within the bandwidth of a single amplifier circuit. The amplifier is now called degenerate.

The degenerate parametric amplifier, needing only two circuits to be coupled to the active element instead of the usual three, is the simplest type to make. It also has the advantage of extra power gain because the total power at the signal frequency is nearly equal to the total power at the idler frequency. Since both lie within the same pass-band both may be detected and the effective gain doubled.

The degenerate amplifier is, however, not invariably chosen because

it has the inherent instability of the lower side-band amplifiers, and it may also not be the best choice in a particular application because of bandwidth and noise figure.

Generalised Manley–Rowe Relations

In the foregoing discussion the power relations quoted have been those relevant to the simple, practical three-frequency parametric or maser amplifier. They have been derived in quantum terms and not using the circuit approach used by Manley and Rowe in their derivation of the generalised expressions:

$$\sum_{m=0}^{\infty}\sum_{n=-\infty}^{\infty}\frac{mP_{m,n}}{mf_1+nf_2}=0$$

$$\sum_{n=0}^{\infty}\sum_{m=-\infty}^{\infty}\frac{nP_{m,n}}{mf_1+nf_2}=0$$

In these expressions m and n are integers, so that the frequencies involved are f_1, f_2, and $mf_1 + nf_2$, where the sign may be positive or negative.

The original paper (p. 62) or a detailed textbook[3] may be consulted for a full derivation which is of considerable mathematical complexity.

[3] *Semiconductor Diode Parametric Amplifiers*, Blackwell and Kotzebue, Prentice-Hall, 1961.

8

Parametric Diode and Ferrite Amplifiers

IN THE PREVIOUS TWO chapters we have discussed systems in which power at one frequency is amplified at the expense of power supplied to the system by a pump at some other frequency. Usually in such amplifiers there is power flow at three distinct frequencies. In the maser-type device these three frequencies are those of transitions between permitted energy levels and power is transferred between them by the mechanisms described in Chapter 6.

In the parametric amplifier coupling between the two circuits carrying power at their separate resonant frequencies is carried out by a time-varying reactance pumped at a third frequency. The reactance may be a *p–n* junction acting as a capacitor and storing energy electrostatically, or a ferrite acting as an inductor and storing energy magnetically.

The variable capacity parametric amplifier will be treated first and in more detail than the ferrite device because it is far more important practically, being in general commercial and military use. The ferrite parametric amplifier, although demonstrated in the laboratory, has been disappointing in practice.

Variable Capacity Junction Diode

The rectifying action of the junction between a *p*- and an *n*-type semiconductor has been discussed earlier (Chapter 3). We are here concerned particularly with the depletion layer formed at the junction. Positive and negative charge is stored in opposite sides of this depletion layer which has insulator properties, so that the *p–n* junction with its stored charge acts like a capacitor. We shall discuss the nature of this capacitor and show that the size of the capacitance depends on the voltage applied across the diode. If this voltage is made to vary with time, then the capacitance will also be time varying and can thus be used as the active element in a parametric amplifier.

The depletion layer is that small region on either side of the junction which no longer contains free-charge carriers because electrons from

the *n*-type material have diffused into the *p*-type, and holes from the *p*-type into the *n*-type material. The *n*-type material is left positively charged and the *p*-type negatively charged by this diffusion process, which continues until a sufficiently great contact potential is established to prevent any further diffusion. It is possible to evaluate the potential difference across the depletion layer and the charge stored, thus obtaining the capacitance.

First let us examine the charge stored by the depletion layer.

In the bulk of the material away from the depletion layer there is no nett charge because the free-charge carriers just neutralise the fixed impurity ions. Thus in the bulk of the *n*-type material there are free

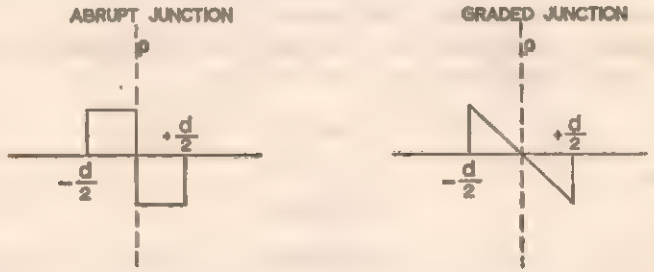

FIG. 8.1

electrons moving at random and a fixed pattern of positively charged impurity ions from which these free electrons originated. In the *p*-type material the free carriers are positively charged and the fixed impurity ions are negative.

If the free-charge carriers are removed, as they are in the depletion layer, then the charge of the fixed impurity ions is no longer neutralised. The exact charge density and distribution will depend upon the way in which the impurity atoms are arranged in the material. If the distribution of impurity atoms in the neighbourhood of the junction is known, then integration over the thickness of the depletion layer will give the charge stored.

The exact distribution of impurity atoms in the junction region depends upon the way in which the diode has been made, e.g. by alloying or by diffusion. In general, however, any diode will approximate either to an abrupt junction of the type described in p. 26, or else to a graded junction where the concentration of impurity atoms changes approximately linearly with distance from *p*-type to *n*-type.

The resultant variation of charge density ρ with distance for a depletion layer of thickness d is shown in Fig. 8.1.

Let us consider unit area of a graded junction where the charge density varies linearly with distance

$$\rho = kx$$

The total charge Q stored on each side of the junction is $\int \rho \, dx$ and thus

$$Q \propto d^2$$

Now Poisson's equation relates electrical potential ϕ and charge density

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{\rho}{\varepsilon}$$

where ε is the permittivity of the material.

Integrating twice gives the total difference of potential over a depletion layer of thickness d as

$$V_0 \propto d^3$$

Thus $Q \propto d^2$

and $V_0 \propto d^3$

i.e. $Q \propto V_0^{\frac{2}{3}}$

The incremental capacitance is given by $C = \frac{dQ}{dV_0}$.

The elastance $S = \frac{1}{C}$.

Thus $C \propto V_0^{-\frac{1}{3}}$ and $S \propto V_0^{\frac{1}{3}}$.

The relation for the abrupt junction calculated in a similar manner is $C \propto V_0^{-\frac{1}{2}}$.

V_0 is the total potential difference across the depletion layer and consists of the contact difference of potential due to the hill in the energy-level diagram, plus any externally applied voltage. The externally applied voltage may consist of a D.C. bias voltage plus any alternating voltage applied from the pump to cause capacity variation.

We are normally only concerned with the time-varying portion of the capacitance caused by the alternating pump voltage at frequency ω_p. For any type of junction the capacity may be written as $C(t)$ indicating a time-varying capacity.

$$C(t) = C_0 + C_1 \cos \omega_p t + C_2 \cos 2\omega_p t + C_3 \cos 3\omega_p t + \ldots$$

Different types of junction will have different values for the coefficients.

In all the previous discussion we have ignored any steady current

flowing through the diode which would constitute a condenser leakage current. The diode must be biased in such a way that such leakage current is very small and in fact the "built-in" bias due to the potential hill allows quite reasonable pumping voltages to be applied without any special external biasing arrangements. This matter is discussed more fully below.

Another matter which has been ignored is the fact that the semiconductor material on either side of the depletion layer will have a small but significant resistance.

These effects can be accounted for by a leakage resistance across the condenser and a resistance in series with it, giving the simple equivalent circuit of Fig. 8.2 for the diode alone.

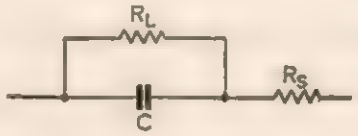

FIG. 8.2

In general, the leakage resistance is sufficiently high to be ignored in comparison with the condenser impedance at normal working frequencies. A more elaborate equivalent circuit including the effects of the diode container is discussed below.

Impedance Matrix for Pumped Varactor

The circuit of Fig. 8.2 strictly represents the varactor diode for a fixed applied voltage and hence a fixed value of C. If the diode is pumped at frequency ω_p, then the capacity varies in the manner described in the last section. There will then be a complex relationship between the currents and voltages in the diode at various frequencies and the simple equivalent circuit will not help in evaluating the properties of any amplifier containing a pumped varactor diode.

A useful tool in the analysis of parametric amplifiers is the impedance matrix representation of the varactor diode. The principles and method of application of the impedance matrix will be outlined below.

The capacity of the pumped varactor has the form

$$C(t) = C_0 + C_1 \cos \omega_p t + C_2 \cos 2\omega_p t + \ldots$$

This may be expressed in the form

$$C(t) = C_0[1 + 2\gamma_1 \cos \omega_p t + 2\gamma_2 \cos 2\omega_p t + \ldots]$$

where $\gamma_1 C_0$ is the half amplitude of the capacitance variation at ω_p and $\gamma_2 C_0$ the half amplitude at $2\omega_p$, etc. The half amplitude is used so that fractional indices are avoided when conversion of the cosine to the exponential form is carried out. It may be met in connection with any periodic quantity, e.g. voltage or current.

$$2\cos\theta = e^{j\theta} + e^{-j\theta}$$

In the impedance matrix it is convenient to use the elastance rather than the capacitance. The relation between the time-varying voltage $V(t)$ applied across the diode and the time-varying current $I(t)$ flowing in it is then given by $V(t) = R_s I(t) + \int S(t)I(t)\,dt$, where $S(t)$ is the time-varying elastance of the pumped varactor and R_s is the series resistance.

Using the exponential form $S(t)$ may be written

$$S(t) = S_0 + S_1 e^{j\omega_p t} + S_1^* e^{-j\omega_p t} + S_2 e^{j2\omega_p t} + S_2^* e^{-j2\omega_p t} + \ldots$$

where the coefficients S_1, S_2, etc., refer to the elastance at the various harmonics of the pump frequency ω_p, and the asterisk indicates the conjugate complex.

If we had been dealing with a very simple linear electrical component in which the current and the voltage were proportional, then we could write $V = ZI$, and we should call Z the impedance. In circuit analysis we should replace the electrical component by a box labelled Z which would tell us all we wanted to know about the current-voltage relationship of the component represented by the box.

Suppose that we now have a more complicated electrical component in which currents and voltages at various different frequencies can exist at the same time. Let us start by considering a simple case in which only two such frequencies occur. A common relationship is then

$$V_1 = a_{11}I_1 + a_{12}I_2$$

$$V_2 = a_{21}I_1 + a_{22}I_2$$

where V_1 and V_2 are the voltages at the two frequencies, I_1 and I_2 the currents, and the a's are coefficients with the dimensions of impedance. Had there been n permitted frequencies, then there would have been n such equations and n^2 coefficients a.

A matrix is a convenient way of writing down such a set of n equations

$$\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{bmatrix}$$

or in the particular case of the two frequencies

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}$$

When the matrix relates currents and voltages in the manner shown above, then it is called an impedance matrix. The theory of matrices[1] develops rapid methods of manipulating matrices in calculation, e.g. the inversion of an impedance to an admittance matrix or the simplification of a large general matrix if certain terms are negligible.

In particular, the previously quoted general equation relating signal currents and voltages in the pumped diode,

$$V(t) = R_s I(t) + \int S(t)I(t)\,dt$$

gives rise to a many-term general-impedance matrix. But in the analysis of a particular circuit this general matrix will be considerably reduced by the recognition of certain simplifying features. Typically, a real problem might undergo gradually more detailed classification of the following type:

(i) Small signals only are involved. The elastance variations are thus due only to the large amplitude pump voltages and are not affected by the signal voltages.

(ii) A particular type of diode is used. The actual voltage–elastance law for this diode is now known, and certain of the elastance coefficients, particularly perhaps those of higher pump harmonics, may be found to be negligible.

(iii) A particular amplifier may be under consideration, e.g. an upper side-band up-converter. Tuned circuits will now restrict the power flow to a limited number of frequencies.

[1] *Determinants and Matrices*, Aitken, Oliver and Boyd.

The impedance matrix in a particular case is thus reduced from the multi-term general form to a simple expression like

$$\begin{bmatrix} V_s \\ V_u \end{bmatrix} = \begin{bmatrix} R_s + \frac{S_0}{j\omega_s} & \frac{S_1^*}{j\omega_s} \\ \frac{S_1}{j\omega_s} & R_s + \frac{S_0}{j\omega_u} \end{bmatrix} \begin{bmatrix} I_s \\ I_u \end{bmatrix}$$

which we shall apply below.

Application of Varactor Impedance Matrix

If the reader is concerned with the calculation of the properties and performance of practical parametric diode circuits, then he should consult a textbook[2] which, although advanced in its field, treats this topic rigorously but from first principles in a manner similar to that used in the more familiar analysis of valve and transistor circuits. Only an indication of the form of such circuit analysis is given below.

Suppose we consider the case of an upper side-band up-converter where the signal power enters at frequency ω_s and the amplified power is extracted at $\omega_u = \omega_p + \omega_s$, where ω_p is the frequency at which the varactor diode is pumped. The incorporated tuned circuits restrict the flow of current to frequencies ω_s and ω_u. The simplified varactor impedance matrix is given by

$$\begin{bmatrix} V_s \\ V_u \end{bmatrix} = \begin{bmatrix} R_s + \frac{S_0}{j\omega_s} & \frac{S_1^*}{j\omega_u} \\ \frac{S_1}{j\omega_s} & R_s + \frac{S_0}{j\omega_u} \end{bmatrix} \begin{bmatrix} I_s \\ I_u \end{bmatrix} \quad . \quad . \quad \text{(A)}$$

where V and I are the half amplitudes of the appropriate voltages and currents. This matrix is then used in the circuit given below (Fig. 8.3).

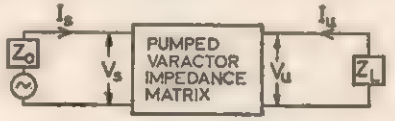

FIG. 8.3

In the circuit Z_0 is the internal impedance of the signal generator which provides an e.m.f. E, and Z_L is the load impedance.

Now from the circuit

$$E = I_s Z_0 + V_s$$

and this relation can be substituted into the matrix equation (A) to eliminate I_s and V_s.

[2] *Varactor Applications*, Penfield and Rafuse, M.I.T. Press.

In fact the matrix, which is really shorthand for two equations, is so simple that the reader might care to write down these two equations and then solve the circuit "longhand".

The result is

$$V_u = \left[R_s + \frac{S_0}{j\omega_u} + \frac{|S_1|^2}{\omega_s \omega_u} \cdot \frac{1}{Z_0 + R_s + \frac{S_0}{j\omega_s}} \right] I_u + \frac{S_1}{j\omega_s} \cdot \frac{E}{Z_0 + R_s + \frac{S_0}{j\omega_s}}$$

This means that the amplifier can be represented by a generator giving an open-circuit e.m.f. equal to the second term, with an internal impedance equal to the term in the first bracket.

The output impedance of the amplifier can now be written down and the output power readily obtained. A similar analysis gives the input impedance and, if noise voltages at the appropriate frequencies are added to the circuit, then its noise properties can be calculated.

The various expressions obtained can be used to discover the best values for those circuit parameters which the designer considers to be especially important in a given application, e.g. there may be a certain load for maximum gain, and a different load for minimum noise figure. They also show how unavoidable features of the real circuits, like R_s the series resistance of the varactor, reduce the amplification and noise performance to something less than the predictions of the Manley–Rowe relations which refer to perfect components.

Practical Parametric Diode Circuits

The foregoing discussion of the varactor diode parametric amplifier has been largely general and theoretical. This section mentions some more practical points like the biasing arrangements, the nature of the frequency-determining circuits, and a typical complete amplifier.

In principle a *p–n* junction diode will act as a capacitor, and the reverse or forward current flowing will be small, for voltages from the breakdown voltage in the reverse direction to very nearly the contact potential, a fraction of a volt, in the forward direction. A steady bias voltage halfway between these two points allows the maximum applied pump voltage. If the diode is not to be fully pumped, then some other bias point may be chosen and in particular, in many cases zero external bias is used. The zero external bias condition is known as self bias because the rectifying action of the diode clamps the bias at a small negative value.

The capsule and leads of a real varactor diode have distributed inductance and capacity which can reasonably be represented by L and C

of Fig. 8.4. The diode thus has a natural resonant frequency and this affects the bandwidth of the amplifier. If the idler frequency is made equal to the diode resonant frequency, then the maximum bandwidth is obtained.[3]

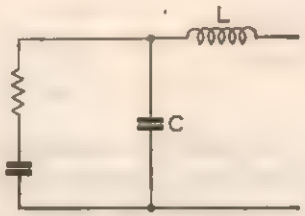

FIG. 8.4

Parametric amplifiers are normally used at frequencies where the frequency selecting circuits can be produced by waveguide or coaxial line techniques rather than by using lumped capacitors and inductors. Such an amplifier operating in the region of 10 Gc/s might have the form shown in Fig. 8.5.

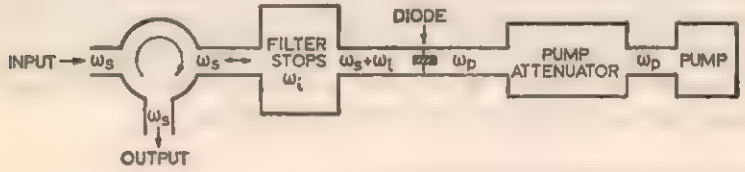

FIG. 8.5

In the system shown the power is carried throughout by waveguides, the frequency of the wave in any particular guide being as indicated. The guide between the diode and the pump attenuator is cut off for ω_s and ω_i, and the pump attenuator controls the amount of pump power reaching the diode and thus acts as a gain control for the amplifier.

The coupling is such that the signal circuit is loaded and hence broad-band while the idler circuit, unloaded except by the diode, largely determines the bandwidth of the amplifier which may be tuned by varying the pump frequency in step with the signal.

The circulator converts the one-port parametric amplifier into a two-port device with input and output separated. The amplifier output is often to a mixer which is noisy and in such a case an extra circulator port terminated in a matched load is inserted between the output and

[3] "Varactor Diode Parametric Amplifiers", Hyde, *Proc. I.E.E.*, Vol. 111, May 1964.

input ports to prevent mixer noise from being added to the original signal input to the amplifier.

Improvements in bandwidth from about 10% to about 50% can be obtained by using a travelling wave parametric amplifier[4] which has about a dozen pumped varactor diodes acting as shunts in a waveguide or transmission line which carries both a signal and an idler wave. Such an amplifier has good noise performance and unilateral gain so that it may be used without a circulator.

Performance and Applications

Varactor diodes parametric amplifiers will operate up to about 100 Gc/s and have noise figures better than any other type of amplifier except the maser. Typical noise temperatures are a few hundred degrees absolute, but this can be improved to about ten or twenty degrees by the use of new junction materials, like gallium arsenide, with liquid helium cooling.[5]

Except at frequencies above about 100 Gc/s, any low noise system can employ a suitable parametric amplifier[6] and current scientific papers and manufacturer's house journals continually refer to new applications in radar, terrestrial and satellite communications, and radio astronomy.

Most operational parametric amplifiers are at present varactor diodes rather than any other sort (see below and Chapter 9), and they are particularly robust and do not necessarily demand refrigeration.

An interesting application of the varactor diode in a non-amplifying role is in the generation of a high harmonic of a high frequency transistor oscillator output so as to provide a microwave source which can replace the expensive small klystron, although the Gunn effect oscillator (p. 29) may prove an even better replacement.

Other Solid State Parametric Amplifiers

In principle any non-linear or time-varying reactance can serve as the coupling element in a parametric amplifier. In particular, an inductor might be used instead of the capacitor treated above.

With the discovery of the ferrite materials there was a possibility of realising a low loss pumped inductor and a microwave parametric amplifier was proposed.[7]

[4] "Investigation of an Experimental Travelling Wave Parametric Amplifier" Mavadoat and Hyde, *Proc. I.E.E. (B)*, Vol. 109, September 1962.

[5] "Operational 4·2° K Parametric Amplifier", Stovman, *P.I.E.E.E.*, Vol. 54, October, 1966

[6] *Semiconductor Diode Parametric Amplifiers*, Blackwell and Kotzebue, Prentice-Hall.

[7] "Proposal for a Ferromagnetic Amplifier in the Microwave Range", Suhl, *Phys., Rev.*, Vol. 106, 1957, p. 384.

The amplifier is pumped at ω_p and the other frequencies ω_s and ω_i are present as in other parametric amplifiers, the active element in this case being a piece of ferrite with a steady magnetising field applied to it, usually in a cavity, although travelling wave versions exist. In general, though the ferrite amplifiers so far produced[8] have disappointed in efficiency and noise performance.

More recently[9] a variable capacitor parametric amplifier has been reported based on the ferroelectric properties of a mixture of barium and strontium titanates.

The complementary ferromagnetic and ferroelectric amplifiers now demonstrated in principle, but hardly comparable operationally with other parametric amplifiers, may well become more important practically as new materials and more efficient methods of excitation become available.

[8] *Coupled Mode and Parametric Electronics*, Louisell, Wiley.
[9] "A Ferroelectric Microwave Parametric Oscillator", Pucel *et al.*, *Proc. I.E.E.E.*, November 1963.

9
Parametric Electron Beam Amplifiers

IN PRINCIPLE THE PUMPING of any variable circuit parameter can lead to amplification. We have already discussed the variable capacitor and its practical realisation in the varactor diode. The variable inductor using ferrite has been made to work as a parametric amplifier, but we have seen that it has disappointed in practice because of difficulty with noise. There is an important and successful class of parametric amplifier where the pumped circuit parameter is not obviously identifiable as a condenser or inductance. This is the electron beam parametric amplifier.

When an A.C. signal is used to set up waves on an electron beam[1] (Appendix III) it is recognised that the beam represents an impedance to the applied signal. Furthermore, this impedance can be caused to vary by the application of a pumping voltage to the beam in a suitable fashion.

Parametric amplifiers have been constructed using electron beams carrying space charge waves of the type used in travelling wave tubes (p. 114). For various reasons, which are discussed later, the beam tubes using space charge waves were unsatisfactory, but a new type of beam wave, the cyclotron wave, was successfully employed. This wave is used in the Adler tube which is a low noise electron beam parametric amplifier now common in communication systems.

It is with the cyclotron wave and the Adler tube that we are principally concerned in this chapter, although other types of amplifier are briefly mentioned later.

Cyclotron Waves—Cuccia Coupler

If an electron moves with a velocity v in a direction perpendicular to a magnetic field of strength B, then the electron will be acted upon by a force Bev perpendicular both to the magnetic field and its own direction of motion. The electron will move in a circular orbit with its plane

[1] The reader who is unfamiliar with electron beam amplifiers of the travelling wave type is advised to look at Appendix III before proceeding with this chapter.

perpendicular to the magnetic field and with radius r, where $mv^2/r = Bev$. The greater the electron velocity the greater the radius of the orbit.

In the cyclotron the charged particle follows a circular path inside an evacuated enclosure consisting of two semicircular metal chambers between which an alternating voltage may be applied. There is a magnetic field perpendicular to the plane of these "dee"-shaped chambers. The voltage applied to the "dees" will give an electric field in the gap between them, which will accelerate or retard the charged particle each time it crosses the gap. If the alternating voltage has the correct frequency, then the particle will be accelerated each time it crosses the gap. The velocity will increase and the next semicircular orbit will be of greater radius, so that the particle will move in a sort of spiral path with continually increasing energy. The frequency of the alternating field for such cyclotron action to occur is given by $\omega_c = \frac{Be}{m}$.

In travelling wave tubes the beam is kept focused by a magnetic field which runs the length of the tube parallel to the electron beam velocity. Thus, if an electron in the beam is given a transverse velocity it will follow a circular path in the plane perpendicular to the beam velocity. The ordinary electron velocity down the tube will now have this perpendicular circular motion superimposed on it and an individual electron will follow a spiral path.

One way of giving an electron a transverse velocity is to apply a transverse electric field by allowing the beam to pass between a pair of deflecting plates. If the voltage applied to these plates alternates at the cyclotron frequency ω_c, then the electron will gain energy continuously while it is in the field, and the radius of the spiral it describes will increase as it passes through the plates.

A pair of plates which impresses this cyclotron modulation on an electron beam is called a Cuccia coupler and the cyclotron frequency is in the hundreds of megacycles region for the sort of magnetic fields normally used for focusing in electron beam tubes.

In Appendix III it is shown that longitudinal velocity modulation of an electron beam can give rise to a fast and a slow space charge wave. A similar investigation of the effect of transverse velocity modulation[2] shows that four waves may be produced on an electron beam.

These are the fast and slow cyclotron waves with velocities given by

$$V = \frac{u_0}{1 \mp \frac{\omega_c}{\omega}}$$

[2] *Coupled Mode and Parametric Electronics*, Louisell, Wiley.

where u_0 is the beam velocity, ω_c the cyclotron frequency, and ω the frequency of the modulating signal.

There are also two other waves, called synchronous waves, each having a phase velocity equal to the beam velocity.

In practical amplifying devices it is the fast cyclotron wave with a forward (+ve) velocity which is used. In this case a snapshot of the beam taken at a particular instant would show it to be a spiral round the axis of the tube. At a later instant the pattern would have moved forward towards the collector end of the tube, its velocity being the fast wave velocity. Each individual electron has a forward velocity u_0 and rotates about the axis at an angular frequency ω_c.

The intersection of the beam with a plane perpendicular to it will describe a circle at angular frequency ω.

As the modulating frequency is allowed to fall to ω_c the spiral opens out and becomes a straight line with infinite wavelength and velocity for the fast wave. We shall examine a little more closely what happens in the Cuccia coupler[3] which launches such a fast cyclotron wave with near infinite velocity.

If power is to be fed on to a beam in the form of a wave, or removed from a beam where it exists as a wave, then the beam must be allowed to interact with a passive circuit which will support a wave of about the same velocity as the beam wave (p. 114). In the ordinary travelling wave tube, where power is to be exchanged with a slow beam wave, a suitable circuit which supports a slow wave, e.g. a helix, has to be used.

Since the fast cyclotron wave is to be used in the particular amplifier we are now considering, then by choosing the signal and cyclotron frequencies to be approximately the same, the wave velocity becomes nearly infinite, and very simple lumped circuits can be used for coupling. Such a circuit, coupling only to the fast wave, is the Cuccia coupler, which is nothing more than two rectangular plates about a centimetre square and about a millimetre apart. All points along the plates change phase together so the circuit wave velocity is effectively infinite. The beam passes between these plates, and an alternating field applied to them gives the electrons transverse velocity and causes them to spiral at the cyclotron frequency. As they pass through the coupler, drawing energy from the alternating field, the radius of the spiral increases and then remains constant after leaving the coupler.

If an electron beam already modulated with a fast cyclotron wave is passed through a Cuccia coupler then energy is extracted from the beam and delivered to a suitable load connected to the coupler. The electrons in the modulated beam will be moving in spirals of large radius when

[3] "The Electron Coupler", Cuccia, *R.C.A. Review*, June 1949.

they enter the coupler, and this radius will decrease as the electrons lose energy in passing through the coupler. If the coupler is matched to the correct load, then all the fast wave energy is removed and the electrons no longer spiral when they emerge from the coupler.

On p. 118 a reference is made to the "stripping" of noise from an electron beam. When a circuit wave is coupled to a fast beam wave there is an interchange of energy between them, the total energy of the system remaining constant. When an electron beam leaves the gun it will carry a certain amount of noise, some of which is fast wave noise which, in a fast wave device, would be added to the noise which was already present on the signal when it arrived at the amplifier input. But when this noisy beam interacts with the signal input circuit the signal power is transferred to the beam, but at the same time the fast wave noise on the beam is transferred to the circuit.

If the input circuit is a properly matched Cuccia coupler, then all the available signal power is transferred to the beam as fast wave energy and all the fast wave noise is removed from the beam.

Adler performed an experiment[4] in which a beam was passed through two Cuccia couplers which he referred to as lumped resonant cavities. The signal was applied to one of the couplers and the amount transferred to a load connected to the other coupler was measured. At the same time the amount of noise appearing at the output was also measured.

If the input circuit was tuned and the magnetic field was adjusted to a certain value, then there was a maximum transfer of signal power from input to output. The same adjustments also gave a minimum noise figure, thus substantiating the idea that the input circuit would couple signal into the beam and noise out of the beam.

Adler suggested that a suitable device interposed between the input and the output couplers could be used to pump the beam and thus produce parametric amplification. In a later paper,[5] with others, he proposed that the quadrupole would be a suitable pumping device.

The Quadrupole

The quadrupole is an arrangement of four electrodes fed with A.C. power at the pump frequency in such a way as to provide a transverse circularly polarised electric field with radial components rotating about the axis at half the pump frequency, the field strength increasing with

[4] "Parametric Amplification of the Fast Electron Wave", Adler, *P.I.R.E.*, Vol. 46, 1958, p. 1300.

[5] "The Quadrupole Amplifier, a Low Noise Parametric Device", Adler *et al.*, *P.I.R.E.*, Vol. 47, 1959, p. 1713.

distance off the axis. A similar field may be produced in a waveguide.

Fig. 9.1 (*a*) shows the electrical connections from the pump to the four electrodes forming the quadrupole cavity. These electrodes should ideally be of hyperbolic section, but circular sections are also used. The plates are tuned with coils, and opposite plates are strapped together to ensure operation in a mode (π mode) which gives the instantaneous polarities shown.

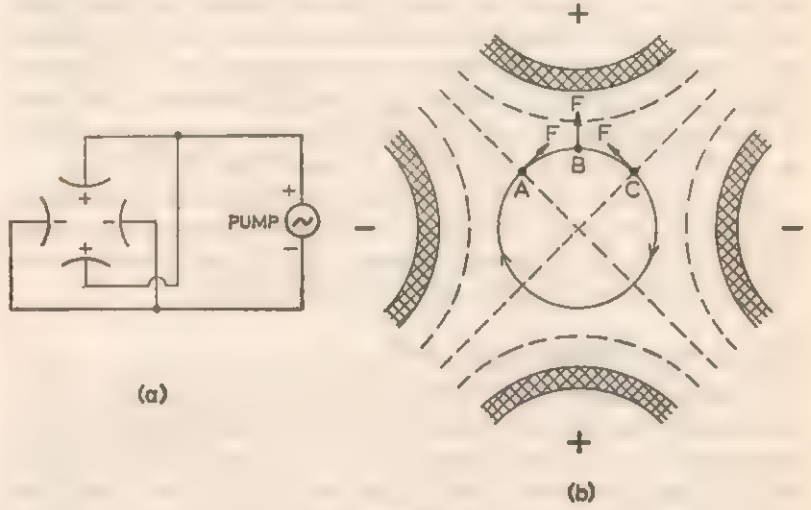

FIG. 9.1

In Fig. 9.1 (*b*) the plates are shown in thick section with instantaneous polarity marked and with equipotential lines shown dotted. The electron beam is passing through the quadrupole in the direction into the paper and the radius of the electron spiral at the section shown is indicated by the circle, the electrons rotating in a clockwise direction.

The directions of the instantaneous forces on three electrons *A*, *B*, and *C* are shown by the arrows labelled *F*. Thus electron *A* will have its orbital velocity increased, *C* will have its orbital velocity decreased, and that of *B* will be unaffected because the force is perpendicular to the orbital velocity.

The quadrupole field and the electrons are both rotating at the same frequency $\omega_c = \frac{\omega_p}{2}$, so that whatever condition (*A*, *B*, or *C*) the electron meets when it enters the quadrupole it continues to be accelerated, unaffected or retarded, as appropriate, all the time it is passing through the quadrupole field. When the electron is accelerated it gains orbital energy from the pump, its radius of gyration increases, and the fast

cyclotron wave energy is increased, conversely the retarded electrons lose energy to the pump and the fast cyclotron wave energy is reduced.

The field due to the quadrupole increases linearly with distance off the axis of the tube. If an electron is accelerated and has its radius of gyration increased when it enters the quadrupole region, then it will move in an orbit of greater radius, encounter a greater quadrupole field and thus have its radius of gyration increased still further. The radius on leaving the quadrupole is equal to the entry radius multiplied by an exponential factor which determines the gain of the tube and which is the same for large and for small amplitude signals, i.e. for large and small radius of gyration on entering the quadrupole region. The radius of gyration of electrons which are retarded will obey exactly the same law except that the exponential index will now have a negative sign.

The gain is calculated by tracing the path of individual electrons as they travel through the quadrupole and then averaging all the possible conditions of phase. On average the exponential growth always exceeds the exponential decay and there is a resultant nett gain.

Adler Tube or Quadrupole Amplifier

The Adler tube using the fast cyclotron wave is the most important of the fast wave parametric amplifiers. Any parametric amplifier may be expected to have a good noise figure, but a simple electron beam usually contains so much noise that parametric amplifiers based upon electron beam tubes generally fail to achieve the expected low noise performance.

The Adler tube has been a practical success for two fundamental reasons. Like any other fast beam wave device it is able to use the fact that fast wave noise can be removed from an electron beam by a suitable passive circuit before the signal is applied to it. But by using the cyclotron wave, which is essentially caused by electron velocities in directions perpendicular to the beam velocity, the Adler tube gets a further advantage, because transverse velocity noise is generated in an electron beam to a lesser extent than the longitudinal velocity noise which affects a space charge wave tube.

There were really three stages in the development of the quadrupole amplifier or Adler tube.

First, Cuccia invented the coupler which was a device for launching power from a generator on to an electron beam as a fast cyclotron wave, or for removing such power from the beam to a load. The electron coupler tube consisted essentially of an input coupler and an output coupler with an electron beam passing through them. The main purpose of Cuccia's tube was to act as a unidirectional connection

between the power source connected to the input coupler and the load connected to the output coupler. It was also possible to modulate this power flow by varying the beam current.

Second, Adler showed that the conditions for launching power from a signal source on to the beam in the input coupler were exactly those which would remove fast wave noise from the beam. Thus maximum power transfer from input to output was achieved with minimum noise figure.

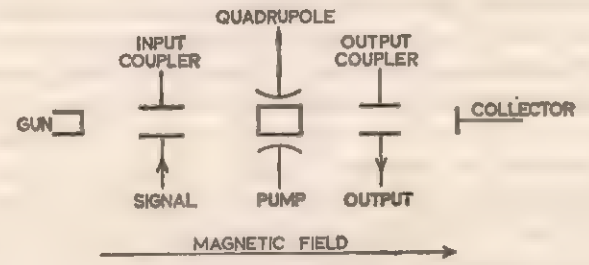

FIG. 9.2

Third, Adler and others showed that a quadrupole device inserted between the input and output couplers could be used to pump the beam and amplify the power at the signal frequency appearing in the load.

Figure 9.2 shows the layout of the important components of an Adler tube. The total tube length might be six or seven inches, with the axial magnetic field provided by a solenoid surrounding the whole tube. The signal input and output, and the pump input, would typically be through coaxial cables.

The details of tube design and construction are covered in such references as 6, 7, 8, and only a few general remarks on design and performance will be given here.

The electron gun will probably be of the multi-anode type with low noise temperature (p. 118) and this temperature may be further reduced by a high magnetic field in the vicinity of the cathode. The collector may be surrounded by electrodes designed to reduce secondary emission which causes extra noise, and the beam voltage will be quite low so that a reasonable number of electron orbits can take place in the length of couplers and quadrupole. With a magnetic field of 200 Gauss a beam voltage of 6 volts gives 4 orbits per cm.

[6] "An Electron-beam Parametric Amplifier for the 200 Mc/s Region", Chalk, *Proc. I.E.E.*, Vol. 108, Part B, No. 37, January 1961.

[7] "A Microwave Adler Tube", Bridges and Ashkin, *P.I.R.E.*, Vol. 48, March 1960.

[8] "Design of Cuccia Couplers for Quadrupole Amplifiers", Chalk, *I.E.E. Electronics Quarterly*, December 1963.

The axial magnetic field determines the value of ω_c and must therefore have the correct value to ensure that this is approximately equal to the signal frequency. The magnetic field is also necessary to keep the electron beam focused. Any interception of the beam by electrodes or other tube structure will cause an increase in noise level. Similarly, if large signals are being handled, the radius of the cyclotron orbits may become so large that collision with electrodes occurs with a similar increase in noise and with saturation effects on the amplifier gain. This last effect will limit the maximum power output of the tube.

The quadrupole may be four curved plates or a cavity, and it is here that the amplification takes place, the power gain taking place entirely at the expense of power from the pump. The gain is independent of the frequency, i.e. the quadrupole itself has an infinite bandwidth.

The Cuccia couplers may also be plates or cavities, the input and output couplers often being turned through 90° relative to each other in order to reduce electromagnetic coupling, and the bandwidth of the amplifier is determined by them. The efficiency with which they couple energy into and out of the beam affects the overall gain of the tube.

The principal features of the performance of the Adler tube are that it is a low noise wide-band amplifier which is unconditionally stable and can be used at frequencies between a few hundred and a few thousand Mc/s.

Fractional bandwidths of about 10% are obtained at a few hundred Mc/s and it has been shown[9] that this figure can be maintained into the microwave region if the current density in the beam is increased as the square of the frequency.

Typically with present tubes the power gain might be 25 db, the noise figure less than 1 db, and the saturated power level about 20 mW.

Other Electron Beam Parametric Amplifiers

Although the Adler tube is the most important practical example, other members of the family of electron beam parametric amplifiers have been constructed.

A transverse modulated electron beam in an axial magnetic field can support four waves (p. 82) and one of them is used in the Adler tube. Direct developments of this tube have been made using the synchronous waves,[10] and the slow and fast waves coupled.[11] This last has been the object of considerable interest because it operates on quasi-parametric principles using a D.C. "pump" which in this case is not the source of power for the amplification process. The noise performance has not, however, been as good as was at first hoped.[12]

There is no theoretical reason why parametric amplification of space charge waves should not be just as successful as parametric amplification of cyclotron waves. If space charge wave parametric amplification is to be attempted, then it is likely to be most profitable from the noise viewpoint if the fast space charge wave is used. This allows the technique of fast wave noise stripping described before to be carried out.

A number of such parametric amplifiers have been described,[13,14] but the performance has not so far been as good as that achieved with cyclotron waves.

[9] *Scaling Laws for Cyclotron-Wave Tubes*, Bridges *et al.*, Munich International Congress on Microwave Tubes, June 1960.

[10] "Amplification of Synchronous Waves", Lucken and Turner, *P.I.E.E.E.*, Vol. 51, September 1963.

[11] "The D.C. Pumped Quadrupole Amplifier—A Wave Analysis", Siegman, *P.I.R.E.*, Vol. 48, 1960.

[12] *Experiments on the Noise Performance of a D.C. Pumped Quadrupole Amplifier*, Vokes and Bridges, NATO–Agard Conference on Low Noise Electronics, Pergamon, 1962.

[13] "Parametric Amplification of Space Charge Waves", Ashkin, *J. App. Phys.*, Vol. 29, December 1958.

[14] "Some Notes on the History of Parametric Transducers", Mumford, *P.I.R.E.*, Vol. 48, May 1960.

Appendix I

Practical Maser Systems

THE FUNDAMENTALS OF MASER action and design have been discussed elsewhere, but little has been said of the way in which this particular type of low noise amplifier is fitted into a microwave system. New features like circulators and refrigeration are needed in maser systems, while some existing components, like duplexers and even waveguide feeders, require to be redesigned and improved.

In the years immediately after the first successful operation of the three-level solid-state maser these amplifiers were incorporated into existing radio systems while still in an experimental state. Now, however, systems are appearing which have been designed with the maser as an integral part. In this section the "hardware" associated with the maser, and some typical special features of such new communication and radar systems are outlined.

Maser Structure, Magnet, Pump

The maser structure itself and the associated magnet and pump are the most obvious novel features of any system, together with the refrigeration which is described later.

The active maser material is usually ruby, which must be inserted into a cavity, or into a waveguide for a travelling wave maser, with circuit arrangements so that pump power can be provided to the ruby and signal power taken to and from it.

A very simple cavity for an X-band maser pumped at K-band, consists of a piece of ruby of the correct size (0·68 × 0·5 × 0·45 in.) with its surface silvered.[1] Two slots cut in the silver layer provide the microwave coupling at signal and pump frequencies and such a cavity can be clamped or soldered to the feeder system.

Two typical feeder arrangements from modern systems[2,3] are shown

[1] "Silvered Ruby Maser Cavity", Cross, *J.A.P.*, Vol. 30, September 1959.
[2] Schimitschek *et al.*, *P.I.E.E.E.*, Vol. 51, February 1963, p. 363.
[3] "Masers in Mars Radar Expt.", Higa and Clauss, *P.I.E.E.E.*, Vol. 51, June 1963.

in Fig. I.1 The system shown in (*b*) employs a coaxial signal feed rather than a waveguide and was one of two masers used in series in the Mars radar experiment.

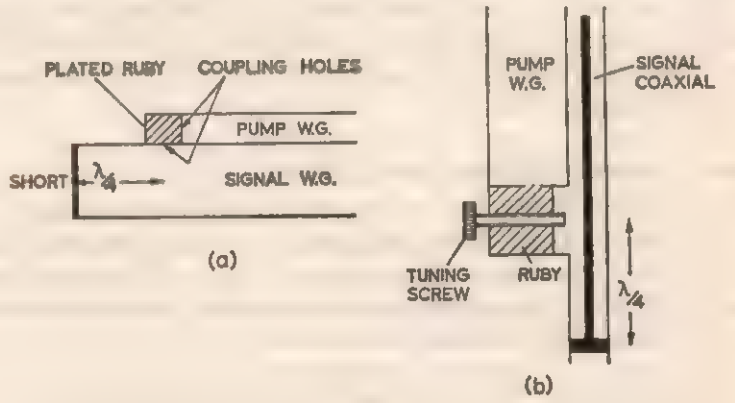

FIG. I.1

In the travelling wave maser, which in general has better bandwidth and power handling capacity than the cavity maser, the ruby material is contained in a waveguide.

A comb structure made of a number of posts (Fig. I.2) slows down the signal wave in the guide so that it is in contact with the active material

FIG. I.2

for a reasonable time. The pump power is propagated down the guide as a fast wave. In the section illustrated, the ruby is mounted on one side of the post, while the ferrite shown absorbs any backward travelling wave and improves the stability.

A typical travelling wave structure, that used in the Telstar communication system,[4] was 5 in. long, and the comb contained 62 fingers loaded on both sides with ruby.

The magnetic field must be at a given angle to the crystal axis of the ruby and its magnitude determines the frequency of operation. The field in the typical travelling wave maser quoted above was 0·33 Wb/m^2

[4] "Masers for Telstar Satellite", Tabor and Sibilia, *Bell Syst. Tech. J.*, Vol. XLII, No. 4, Pt. 3, July 1963.

provided by a permanent magnet which was kept at a constant temperature because the field variation was 4×10^{-5} Wb/m² per degree Centigrade, and the tuning variation was 2400 Mc/s per Wb/m².

The pump was a reflex klystron with an AFC system which kept its frequency within a few Mc/s of 30,180 Mc/s. It delivered about 70 mW of power since a fall in pump power below 50 mW would reduce the maser gain by 2 db.

Noise Temperature Due to System Elements

Circuit elements which are intrinsically noisy like mixers, and lossy components like feeders and duplexers, all increase the total noise temperature of the system. The reduction of such effects may require a new fundamental principle in the system design, or merely piecemeal improvement in the individual components.

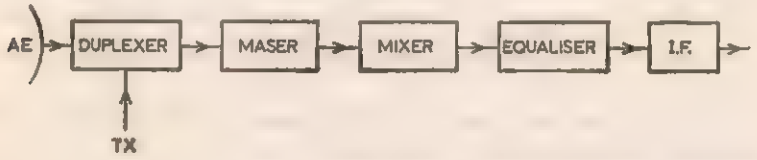

FIG. I.3

As an example of the low noise requirement radically affecting the system layout, the R.F. stage of the Telstar receiver system is worth considering.

The chosen maser at liquid helium temperature had 42 db gain with 16 Mc/s bandwidth between 3 db points, while at 25 Mc/s bandwidth the gain had fallen to 35 db. The system was, however, required to have a bandwidth between 3 db points of 25 Mc/s, and this could have been achieved by stagger-tuning the maser.

The operating frequency of any maser is determined by the applied magnetic field. If the first half of the structure of a travelling wave maser is in a different field from the second half, then the two halves will tune to different frequencies and staggered tuning will result. A bandwidth of 25 Mc/s can be obtained but the gain is only 27 db, which would be adequate if it were not for the very large noise temperature of the mixer stage which follows the maser.

This mixer has a 12 db noise figure which corresponds to a noise temperature of $(15{\cdot}9 - 1)290 = 4321°$ K.

If the maser gain is 27 db (500)—the stagger-tuned value—then the contribution of the mixer to the noise temperature at the maser input terminals is $4321/500 = 8{\cdot}6°$ K. This would greatly degrade the noise

temperature at the maser input because the figure due to the maser and its input waveguide alone is only 3·5° K.

With the single-tuned maser the gain is 42 db (15,894) at mid-band, and 35 db (3162) at 25 Mc/s bandwidth.

Thus the noise temperature contributed by the mixer at the maser input terminals is $4321/15{,}894 = 0{\cdot}27°$ K at mid-band, and $4321/3162 = 1{\cdot}36°$ K at the band edge.

The receiver response was now given the required shape by the equaliser circuit, a bridged T filter inserted after the mixer, which gave a 7 db loss at mid-band and 3 db loss at 25 Mc/s bandwidth. The overall gain thus became $42 - 7 = 35$ db at mid-band and $35 - 3 = 32$ db at the band edge.

In contrast to the foregoing example of a fundamental modification to system layout consequent on the use of maser amplification, the first successful use of a maser in a conventional X-band radar, of peak pulse power 150 kW, necessitated the improvement of one of the components—the duplexing system.[5]

The conventional TR switch allowed a few milliwatts of transmitter power to break through to the maser. This reduced the population in the higher level and caused the maser gain to fall. In addition, the breakthrough power varied from pulse to pulse and the fluctuation in receiver performance was unacceptable. An extra 30 db of isolation was required to protect the maser from saturation.

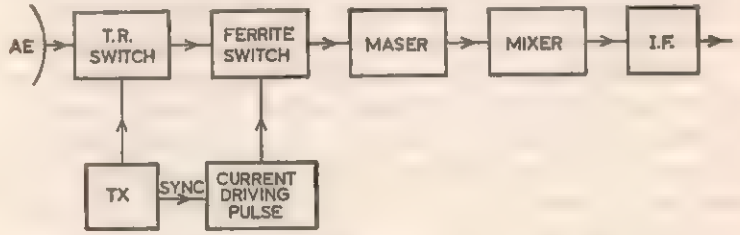

FIG. I.4

A high-speed ferrite switch, inserted between the TR switch and the maser, provided 30 db of isolation during transmission, but only 0·25 db of loss, corresponding to $0{\cdot}06 \times 290 \simeq 17°$ K increase in the system noise temperature.

The switch consisted of a tapered ferrite rod in the centre of a waveguide with conducting fins on each side of the rod. With no D.C.

[5] Goodwin, *P.I.R.E.*, Vol. 48, January 1960, p. 113.

magnetic field applied the ferrite acts as a dielectric and the fins are in an equipotential plane and cause no attenuation of the normal mode. During the transmitted pulses (2·35 microseconds long at 416 p.r.f.) a current pulse produces a longitudinal magnetic field which is applied to the ferrite. The wave is no longer propagated unaffected through the structure (Appendix IV) and more than 30 db of attenuation is obtained over a bandwidth of 120 Mc/s.

The noise temperature of the whole system was 173° K, which compared well with the figure of about 2000° K for a good X-band radar with no maser.

Refrigeration

Perhaps the greatest novelty for the electronic engineer in a low noise microwave system is the refrigeration, although those who have worked with infra-red detectors will already have encountered it.

In general, liquefied nitrogen (77° K) or helium (4·2° K) in a dewar vessel provides the low temperature environment. Because of its cheapness compared with helium, nitrogen is used when possible, either as the only cooling if the system does not demand the very lowest temperatures, or as a precooling liquid in a double dewar or some other arrangement, to reduce the loss of liquid helium which is used to provide the final cooling.

In a maser system the low temperature is provided for two reasons: first the maser will not work at all unless its temperature is well below room temperature (p. 57), and second, the degradation in noise performance due to lossy elements can be much reduced if they are cooled.

Typically an element like a circulator occurring in a system before the first amplifying stage will, if lossy, increase the noise temperature of the system considerably. An element at room temperature with 0·5 db loss has an absorption coefficient of 0·11 and will increase the noise temperature by about 32° K, while at helium temperature the degradation will be less than a degree.

It is therefore worth while to design a helium-cooled circulator to operate with a travelling wave maser.[6] One of the main considerations in this design is thermal performance so that loss of liquid helium is not too great. This particular circulator dissipated 35 mW of heat, made up of joule heating in the coils generating the magnetic field for the ferrite circulator, and eddy current losses due to switching of this field.

Quite frequently, where the introduction of a maser has necessitated helium cooling, thermodynamics plays an important part in the system performance.

[6] de Gruyl *et al.*, *P.I.E.E.E.*, Vol. 51, June 1963, p. 947.

For instance, a travelling wave maser may be able to tolerate, without disturbance to its output, one or two watts of power outside its pass band breaking through to its input terminals from a neighbouring transmitter. But the dissipation of so much power may involve an unacceptably large loss of liquid helium, so that better isolation reducing the total breakthrough at all frequencies to not more than a few hundred milliwatts is required. The isolation at the signal frequency must, of course, be much better in order to avoid saturation.

The maser input leads have always made the largest contribution to the noise temperature of the maser amplifier itself and their design involves conflicting electrical and thermal requirements. Thin wall, stainless steel coaxial cables reduce heat conduction, and thin copper plating improves electrical properties, a compromise being reached between the relative amounts of steel and copper employed.

In the Telstar maser (p. 92) waveguide input leads were used instead of coaxial to get the lowest losses. The waveguide was 0·020 in. thick seamless stainless steel internally plated with 0·0002 in. copper. The room temperature loss of the maser feed with this guide was about 0·1 db. A number of such guides were made and selected for the best noise performance.

Some typical performance figures of the Telstar system are of interest in terms of the application problems discussed above:

Centre frequency	4170 Mc/s
Effective bandwidth	25 Mc/s
Effective gain	34·5 db (equalised)
Pump frequency	30,175 Mc/s
Pump power	70 mW
Magnetic field	0·33 Wb/m²
Overall maser noise temperature	3·5° K
Bath temperature	4·2° K
Liquid helium consumption	½ litre/hour
Helium capacity	10 litres

Future Development

It seems likely that future development in maser systems may lie in three main fields: the maser itself, the associated radio equipment, and the cryogenic system.

In the maser itself materials, pumping, and operating techniques are likely to be examined with a view to improving the gain-bandwidth product at a given temperature, raising the operating temperature,

improving the efficiency, and extending the frequency range particularly in the higher direction.

Optical pumping in a ruby maser was first demonstrated some years ago[7] and more recent work[8,9] has shown that optically pumped microwave masers can be operated at much above 4° K, typically at liquid nitrogen temperature and higher, without the deterioration in noise temperature and gain-bandwidth product which occurs with microwave pumping at elevated temperatures. In addition, the use of optical pumping allows the maser frequency to be extended beyond the range of present devices using microwave pumping sources.

We have already discussed some of the ways of reducing the degradation in noise temperature due to the associated circuit elements like feeders, duplexers, switches, etc. The figures quoted below on the Telstar system noise temperature, together with those quoted earlier (p. 9) on the Echo project, give some idea of where there is scope for improvement, and of those contributions like the 2·4° K due to the atmosphere, which cannot be avoided—except by going outside the atmosphere into space.

The contributions to the total noise temperature with the aerial at 90° elevation are:

Absorption in dry atmosphere		2·4° K
Aerial side lobes	1·0	
Feeder circuits, etc.	14·2	
Maser	3·5	19·2° K
Second stage	0·5	
Radome absorption (dry)		3° K
Radome scattering (dry) (ground noise scattered into aerial horn)		7·4° K
		32° K

The noise temperature increases to 42° K at 7·5° elevation due to greater path in the absorbing atmosphere. It rises very sharply to over 200° K when the aerial beam strikes the ground. A wet atmosphere also causes an increase, and a wet radome may put the temperature up by over 20° K.

Perhaps most progress is likely to occur in the near future in the cryogenic techniques. Closed-circuit helium liquefiers are now being

[7] Devor *et al.*, *Phys. Rev., Letters 3*, November 1959, p. 468.
[8] Hsu and Tittel, *P.I.E.E.E.*, Vol. 51, January 1963, p. 185.
[9] Szabo, *P.I.E.E.E.*, Vol. 51, July 1963, p. 1037.

developed and some thousands of hours experience have been acquired in operating masers in such systems.[10] Typically, a system will consist of a compressor and a refrigerator unit which may be separated by as much as 200 ft. Ten days of operation of the closed-circuit system without adjustment has been reported with the refrigerator unit mounted on the aerial of a low noise system.

[10] Higa, Wiebe, *P.I.E.E.E.*, Vol. 51, May 1963, p. 851.

Appendix II

Laser Systems and Applications

THE GENERAL REQUIREMENTS FOR the achievement of a non-equilibrium system at optical frequencies have been discussed elsewhere (p. 48 and p. 57). The purpose of this appendix is to describe some typical laser systems with their associated equipment and to discuss some of the applications of the laser.

The laser is almost always met as a radiation source, i.e. as an oscillator rather than as an amplifier. The amplifier is therefore not discussed until the end of this appendix and we must start by considering how an active laser material can be made into an oscillator.

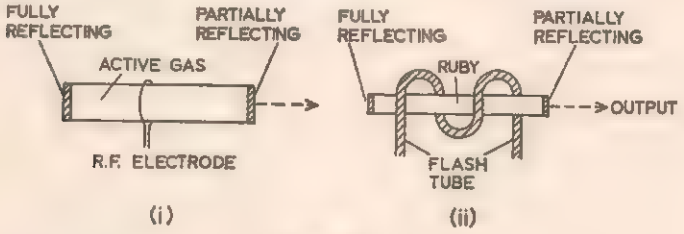

FIG. II.1

The general requirements in any oscillator are power gain, provided in the laser by the active material, feedback of sufficient power to maintain oscillation, and a method of radiating the remainder of the power externally. We shall first consider the sort of system which provides the required feedback and output.

In general, the active laser material is mounted between two parallel mirrors as shown in Fig. II.1 (i) and (ii). In the case of the gas laser[1] the mirrors form the ends of the tube in which the discharge is maintained by an external coil or coils connected to a radio frequency oscillator. In the solid laser the active material, e.g. ruby, in the form of a rod has optical flats ground on the ends and these are coated to give the required reflection. The ruby is pumped to the active state by the light output

[1] "Gas Lasers", Bloom. *P.I.E.E.E.*, Vol. 54, October 1966.

from a flash tube, containing for example xenon, which is coiled round the rod.

One of the mirrors is made totally reflecting and the other partially reflecting, so that enough energy is fed back through the active material to maintain oscillation. The light escaping through the partially reflecting mirror is the laser output. The space between the mirrors is, in fact, a resonant cavity many wavelengths long, whereas in the microwave maser the active material is often mounted in a cavity with dimensions of the order of only one wavelength. The whole of the cross-section of the partially reflecting mirror constitutes the source of light, and it should be filled with radiation of one frequency, phase, and intensity. This is generally the case with a gas laser, but the solid laser tends to give a number of spots on the end of the rod rather than uniform illumination. The light from different spots will be of different frequency, phase, and intensity, each spot representing the output due to one particular mode of oscillation of the long optical resonator. Inhomogeneities in the crystal encourage the occurrence of filamentary laser action along separate paths in the rod, and since the conditions for oscillation vary slightly with temperature, the general character of the laser output tends to change during the time of operation. Much of the energy provided by the flash tube tends to heat the ruby and the dissipation of this heat is a difficulty. Consequently most, but not all,[2] solid lasers operate under pulsed conditions, so that the temperature rise is not too great. Furthermore, the flash tube is required to produce several hundred joules of pumping energy in order to maintain laser action, and this is normally obtained by discharging a bank of condensers through the lamp under pulsed conditions. Less than 1% of the energy supplied to the ruby appears as coherent light, and without elaborate precautions[3] there is a tendency for the output to be spiky rather than a clean, square pulse of light.

Both the gas and the solid laser show this very low efficiency whereas that of the injection laser (p. 129) is very much greater. In general, the gas laser provides better coherence and easy continuous wave action, while the solid laser can provide very high peak power output of the order of megawatts.

Modified Laser Oscillator Systems

The plane parallel mirrors placed at either end of the active laser medium constitute a Fabry–Perot interferometer system. It has been

[2] "A Continuously Operating Ruby Optical Maser", Nelson and Boyle, *Applied Optics Supplement 1, Optical Masers*, 1962.
[3] *Lasers*, Lengyel, Wiley.

shown[4] that much greater separation of the possible resonator modes will occur, and much less power loss due to diffraction of the parallel beam at the mirrors, if a confocal mirror system is employed instead of the plane parallel arrangement. Two spherical mirrors are set up so that the centre of curvature of each lies in the surface of the other and their foci coincide in the centre of the active material. These mirrors are now outside and quite separate from the ruby rod or the gas discharge tube. In the latter case a considerable advantage is gained because the comparatively delicate highly reflecting mirror coatings no longer suffer the baking or other outgassing to which the discharge tube must be subjected. One of the principal advantages of the confocal mirror system is that the angular tolerances for oscillation to occur are much less stringent than in the plane parallel case and the laser is very much easier to set up.

When the plane mirrors are removed from the ends of the discharge tube in the confocal system then the tube must be closed by windows which contain the gas but allow the multiple passage of a light beam with the minimum loss due to absorption or reflection. This is commonly done by sealing the ends of the tube with optical flats made of a suitable low absorption glass and set at the Brewster angle to minimise reflection.

The problem of getting the laser light out of the gas laser has its complement in the problem of getting the maximum amount of pump light into the ruby. A composite arrangement using sapphire, which has a high refractive index like ruby, is often used (Al_2O_3 with Cr impurity is ruby, without Cr it is sapphire). In the arrangement shown in Fig. II.2 pump light impinging on the wide end of the sapphire trumpet is

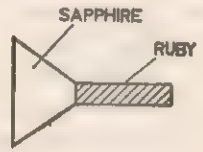

FIG. II.2

all fed into the active ruby. As well as providing a way of getting pump light into the ruby, the sapphire in a composite arrangement provides a way for heat to escape.

If a pulse of optical pump power is applied to a laser, then the number of elements in the high energy state starts to increase at the

[4] "Resonant Modes in a Maser Interferometer", Fox and Li, *B.S.T.J.*, Vol. 40, 1961.

beginning of the pulse and builds up until population inversion is just achieved, laser action occurs and the upper energy level becomes depopulated. If the pumping pulse is still being applied when the laser action is in progress then the population of the upper level might conceivably be reduced below the threshold required for oscillation and raised to it again once, or even several times, in the duration of the pumping pulse. The laser output pulse would then be expected to be spiky in character and this is certainly the case with many pulsed ruby lasers. A better-shaped pulse, and in particular a higher peak pulse power, can be obtained if a population inversion much greater than the threshold required for oscillation is established by pumping before laser action is allowed to start.[5] One of the mirrors could be removed or obscured by a shutter until a sufficiently high degree of inversion were achieved, or some attenuating material could be introduced into the optical path to reduce or "spoil" the regenerative properties, or "Q", of the optical cavity until the laser was required to fire.[6] A wide variety of different devices such as rotating mirrors, electrically operated Kerr cells, and mirror coatings with controlled reflection coefficients have been used to bring the cavity to the regenerative condition at the right moment.

Practical Laser Systems

To show the sort of apparatus used and the way it is laid out in practical lasers, examples of the gas, the solid, and the injection laser are briefly described below. In each case an attempt has been made to select a system which is typical of its class, but which incorporates at least something to minimise the defects to which its class is subject, e.g. we shall consider an array which reduces some of the disadvantages of the small emitting area of the injection laser.

Figure II.3 shows the sort of apparatus that would be used in a typical CW gas laser. A quartz or silica discharge tube, about a centimetre or two in diameter and something less than a metre long, contains a mixture of helium and neon in proportions of about five to one with a total pressure of the order of 0·5 mm Hg. The helium–neon mixture will give laser action at about a dozen frequencies in the visible and infra-red regions. A given composition and pressure seems to be best for any particular frequency, which will be largely determined by the selective nature of the reflecting coating on the mirrors. The discharge tube is

[5] "Q-switched Optical Masers", Midwinter and Forrester, *I.E.E. Laser Symposium*, September 1964.

[6] "Giant Optical Pulsations from Ruby", McClung and Hellwarth, *Applied Optics, Supplement 1, Optical Masers*, 1962.

well baked, and precautions are taken against impurities when filling it. The ends of the tube are closed by optical flats set at the Brewster angle and the light emerging and that taking part in the laser action will consequently be plane polarised. The RF discharge is maintained in the gas by two external electrodes wrapped round the tube and connected to an oscillator giving about 20 watts of power at about 30 Mc/s. The energy in the discharge must be sufficiently great to allow good pumping, but too much energy reduces the life of the tube which is typically several hundred hours.

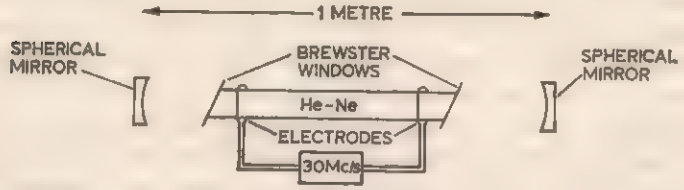

FIG. II.3

The external spherical confocal mirrors have a multilayer dielectric coating of the appropriate material and thickness to give high reflection at the desired wavelength. These reflecting surfaces need to be protected as far as possible from dirt and damage, and they will have a useful life comparable with that of the discharge tube. One of the mirrors will be arranged to transmit a small fraction of the light and this is the laser output which may be about 100 mW.

The light emitting area in an injection laser is an extremely small strip on the face of the junction diode. If a much larger area is required, in order to increase the total power in the beam or to reduce its diffraction beamwidth, then the necessary increase in junction size—particularly the junction length—would give rise to great difficulties because of the large driving current needed and because of the complications in the collimating optics.

An injection laser array has been produced[7] where the drive current is passed in series through a number of diodes which each has its own collimating lens. In Fig. II.4 the diagram is a section, the light emitting junction being virtually a line and the lens cylindrical. With ten laser diodes, mounted on copper blocks and connected thermally to a heat sink cooled with liquid nitrogen, 400-watt peak light output pulses were

[7] "A Semiconductor Laser Array", Broom, *I.E.E. Laser Symposium*, September 1964.

obtained with 200-ampere peak driving current pulses of length 0·7 microsecond.

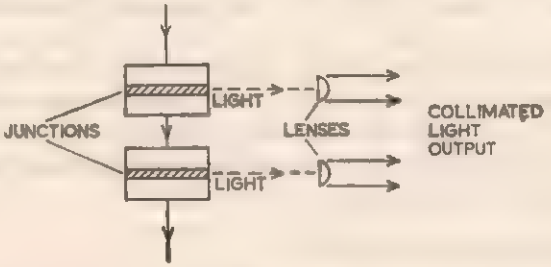

FIG. II.4

In the case of the ruby or other solid laser[8] there are really two classes of practical system to consider: there is the system designed to give CW operation[9] and the "giant pulse" system which aims at the largest possible peak pulse output. In each case there is a general requirement for a high quality crystal laser rod and good mirrors. But in the CW case there is concentration on continuously operating high energy pump lamps and techniques, like sapphire sheathing for the active rod and liquid nitrogen cooling, to get the most efficient use of pump energy[10] and the best removal of heat. In the giant pulse systems[11] the laser action is held off until the highest state of population inversion is achieved and much interest is concentrated on the various types of "Q spoiling" like the spinning prism, the Kerr cell, or the synchroinsed ultrasonic shutter.[12]

Applications of Laser Action

The applications of the laser tend to occur at three levels of increasing sophistication. There is, first of all, the ability to concentrate very high power on to a very small spot; this may be used in machining or certain research where very high local temperatures are required. Second in sophistication is the use of the fact that the high power is at a single well-defined frequency; this may be used in more selective atomic or molecular excitation processes, particularly in fundamental research but

[8] "Crystalline Solid Lasers", Kiss and Pressley, *P.I.E.E.E.*, Vol. 54, October 1966.
[9] "Excitation, Relaxation, and Continuous Maser Action in Calcium Fluoride", Boyd *et al.*, *Phys. Rev., Letters*, Vol. 8, 1962.
[10] "Pumping Power Considerations in an Optical Maser", Svelto, *Applied Optics, Supplement I, Optical Masers*, 1962.
[11] "Giant Optical Pulsations from Ruby", McClung and Hellwarth, *J.A.P.*, Vol. 33, 1962.
[12] "Ultrasonic Refraction Shutter for Optical Maser Oscillator", De Maria *et al.*, *J.A.P.*, Vol. 34, March 1963.

also in the optical pumping of masers. Most advanced of all is the use of the fact that the high power is at a very high frequency and hence can provide an enormous bandwidth when modulated in a communication system.

These applications are briefly discussed below with references to papers giving more detailed descriptions. Developments are, however, proceeding so rapidly that the current journals will almost always contain papers of importance which make out of date the performance figures published in textbooks.

Laser as Cutting or Welding Tool

One of the most frequently exhibited laser demonstrations is the punching of a hole in a razor blade by a focused laser pulse which may have a power density of 10^4 or 10^5 Mw/cm^2. in a spot of diameter 50 microns or less. The possibilities of such a precise tool in micro-machining are fairly obvious, but in such an application the regular pulses at fairly high repetition rate available from a pulsed gas laser are more likely to be used than single giant pulses from a spoiled Q solid laser.

For machining, the material must be removed by melting and subsequent vaporisation, but these processes when brought about by a laser pulse are of considerable complexity and much must be discovered about them before machining is widely carried out with the laser.[13] At a lower power level melting alone may be achieved and this has been used in welding processes in such applications as microminiature electronic circuits. The laser tool has perhaps been most developed in the medical field, particularly in eye surgery[14] and in fairly conventional radiotherapy of the X-ray type.

Fundamental Investigations

The laser is a very high-powered source of radiation at a single optical frequency, or at any rate spread over a very narrow frequency band. When focused, the power available per unit area of target is higher by many orders of magnitude than ever before known and most applications of the laser in fundamental science exploit this previously unobtainable level of power. In the most direct way, the power can be used to produce very high temperatures where normal heating is inconvenient and work of this sort has been carried out on high

[13] "Vaporisation by Laser Beams", Hughes, *I.E.E. Laser Symposium*, September 1964.

[14] "Use of Ruby Laser for Retinal Photo-coagulation". Smart, *I.E.E. Laser Symposium*, September 1964.

temperature plasma.[15] Such very high energy environments, with electromagnetic fields at well defined frequencies, may also be important in the investigation of chemical reactions.

There is a number of optical processes in which the frequency of the light involved is changed. Research in this field is very important because it is expected to yield information about fundamental radiation processes and about the structure of matter. In these processes the amount of light suffering such a change of frequency is very small and the experimental work has been greatly restricted by the difficulty of detecting it. Indeed, the theory of some of the most interesting phenomena which had been predicted, showed that there was no possibility with existing instruments of detecting the small amounts of light involved. With the large laser power outputs available, detection is simplified in those cases like Raman scattering where the effect had already been observed, and made possible for the first time in the case of the non-linear processes.

Raman effect studies once involved long photographic exposure times to detect the scattered light, but the experiments are now greatly aided and extended by the monochromatic light of high power available from the laser[16]. The Raman scattered light from organic liquids, which is monochromatic but different in frequency from the original laser light, may in itself constitute a useful radiation source and more recently it has been shown[17] that stimulated emission can occur in suitable Raman sources.

The theoretical prediction of multiquantum effects[18] in certain crystals has been verified using laser outputs.[19] Typically two photons may be absorbed simultaneously to raise the material from a low to a higher energy level. A subsequent downward transition between the same levels may occur with the emission of a single photon at twice the frequency of the two absorbed photons. Practical applications in harmonic generation and mixing are likely to be very important.[20]

Communication and Ranging Systems

The frequency of the radiation from a laser is of the order of a million times that used in communication systems at present and, in theory at any rate, a laser communication link could carry a million times more

[15] Minck, *J. App. Phys.*, Vol. 35, 1964.

[16] "Ruby Optical Maser as a Raman Source", Porto and Wood, *J. Opt. Soc. Am.*, Vol. 55, 1962.

[17] Woodbury, *Proc. I.R.E.*, Vol. 50, 1962.

[18] Loudon, *Proc. Phys. Soc.*, Vol. 80, 1962.

[19] Giordmaine, *Phys. Rev., Letters*, 8, 1962.

[20] Bass *et al.*, *Phys. Rev., Letters*, 8, 1962.

channels than a conventional one. Because the wavelength is so small the divergency of the laser beam will be very small indeed, and a reflector of the same sort of size as a microwave aerial would, again in theory, give extremely high directivity and corresponding sensitivity, which would imply long detection ranges in optical radar and communication systems even with quite low-powered sources.

It is as well to reconsider the enormous theoretical advantages of optical communication systems in more practical terms at an early stage. Although such considerations may modify early enthusiasm for laser systems, it is worth remembering that the failure to achieve the theoretical performance figures is because of inadequacies in "hardware" which is in the very earliest stages of development and may show great improvements soon.

Although the short wavelength of the optical radiation means an extremely narrow beam from a radar size reflector, there is an implication that such a reflector can be made to optical tolerances and that it will retain its shape and surface qualities in use. Furthermore, although a very narrow beam will give high sensitivity, it will be extremely difficult to hold on the target and some auxiliary tracking system may well be needed.

The full realisation of the very large bandwidth theoretically possible with a laser system depends upon certain practical considerations. A suitable efficient method of modulating the light beam must be available and such modulators are only just being developed. They are usually electro-optical devices[21] like Kerr cells operating on the light beam for ordinary lasers, and simpler electrical modulators operating directly on the driving current in the injection laser. Even if the modulation is efficient, the frequency stability of present lasers, particularly those of high power, will have to be much improved if many closely spaced communication channels are to be handled on the same carrier. In addition, a superhet receiver must be developed if the possibilities of the coherent system are to be fully exploited.[22]

Assuming that suitable terminal equipment is designed for an operational laser communication system there are still problems associated with propagation through the intervening medium. At optical and near infra-red frequencies the attenuation in a clear atmosphere is fairly great, and under foul weather conditions it will rise considerably. The choice of operating frequency must be made with

[21] "Electro-optical Modulation at Microwave Frequencies", Harvey, *I.E.E. Laser Symposium*, September 1964.

[22] "Superheterodyne Reception at Optical Frequencies", Warner and Warden, *I.E.E. Laser Symposium*, September 1964.

care to avoid peaks in the absorption spectrum, and even if the signal intensity is adequate, there may well be phase and path variations due to turbulence which will degrade the system performance.[23] Since commercial laser communication systems would presumably be of very large capacity it may well be economic to use guided propagation in glass fibre pipes or cables.[24, 25]

Pilot models of laser communication equipment have been built using ordinary and injection lasers, while useful range-finding equipment for short-range military tasks and for long-range space applications also exist. Progress is rapid and the current journals or the report of the latest laser conference will provide the best information on new applications.

[23] "Unguided Optical Propagation in the Atmosphere and Undersea", Meredith, *I.E.E. Laser Symposium*, September 1964.

[24] "Communication Systems in the Visible and Infra-red Spectra: Present and Future", Karbowiak, *I.E.E. Laser Symposium*, September 1964.

[25] "Optical Transmission Research", Miller and Tillotson, *P.I.E.E.E.*, Vol. 54, October 1966.

Appendix III

Electron Beam Amplifiers

IN MANY ELECTRONIC DEVICES, including the conventional valve, amplification is achieved by converting some of the D.C. energy of an electron stream into A.C. energy in a suitable circuit. In the following section we are concerned with long well-focused beams used in the Gc/s region in such devices as the klystron and the travelling wave tube.

Attempts to understand electron beam tubes in terms of the ballistics of single electrons are much complicated by the effects of space charge. If the electron beam is considered as an elastic medium which resists displacement of the electrons from their uniform density, then useful information about these devices can be obtained in terms of the wave-like behaviour of compressions and rarefactions in the electron beam.

Space Charge Waves

Let us first reduce the problem to a single dimension by considering the simplest possible beam consisting of electrons moving in single file through a perfect vacuum. Originally the electrons all move at the same velocity, but modulation may be imposed on the beam by passing it for a short part of its path through a small A.C. electric field. Successive groups of electrons passing through the field will be retarded, unaffected, and accelerated as the electrostatic force due to the alternating field changes from a maximum in the direction opposite to the electron velocity, through zero, to a maximum in the direction of the beam velocity.

After leaving the field the electrons will have different velocities. Those retarded will have low velocities and will fall back towards the unaffected electrons immediately behind them, while these unaffected electrons will also be approached from the other side by the accelerated electrons now moving with high velocities. Some distance away from the alternating field the electron beam, originally of uniform density, will consist of high density bunches separated by low density gaps, the

centre of each bunch and gap being an unaffected electron which passed through the A.C. field when its value was zero.

If this simple case is considered in detail it can be shown, e.g. by using an Applegate diagram,[1] that optimum bunching occurs at some point a certain distance from the alternating field. Many of the electrons which passed in succession through this field will arrive at the optimum bunching point simultaneously.

The effect of the velocity modulation imposed by the alternating field is to produce bunches and gaps, i.e. the electron beam, originally of uniform density, is compressed and rarefied. The electrostatic force of repulsion between electrons will, however, tend to oppose such compression and rarefaction and will act like the restoring force which resists displacement in an elastic medium.

Any disturbing force tending to modulate the uniform beam is opposed by the electrostatic force tending to reduce to zero the displacement of electrons from their equilibrium positions. The result of any disturbance will be an oscillation in which the electron displacement varies periodically in a manner determined by the general properties of the electron beam.

It is fairly obvious that there is a possibility of wavelike behaviour if the electron beam is disturbed. We shall not investigate this idealised case any further but shall consider similar oscillations in the more complex beams of electrons and positive ions which are used in practical devices. These oscillations are called plasma oscillations.

Plasma

A plasma is a region where there is a large concentration of ions equally, or nearly equally, divided between positive and negative so that the space charge is approximately neutral. Un-ionised molecules will also usually be present in a plasma, which may occur in a gas, a liquid, or a semiconducting solid.

We shall be concerned with the gaseous plasma which has been very extensively investigated because of its occurrence in many fields of great theoretical and practical importance, such as the gas discharge, the ionosphere, the thermonuclear fusion reactor, and the microwave beam amplifiers which are our particular concern.

There is sufficient residual gas in a "hard" vacuum tube for a beam of fast-moving electrons to produce considerable numbers of positive ions. The positive gas ions tend to move along the axis in the low potential region due to the electrons, although some escape from the

[1] *Electron Physics and Technology*, Thomson and Callick, E.U.P., 1959.

beam and reach the walls of the tube where they are neutralised. In general, the rate of production of positive ions is sufficiently high for the space charge in the beam to be neutralised enough to enable it to be treated as a plasma.

For an idealised beam which is completely neutralised, of infinite cross-section, and with small axial variations in charge density due to oscillation, it is possible to calculate the natural radian frequency (ω_p) of plasma oscillation.[2]

$$\omega_p = \sqrt{\frac{e}{m}\frac{\rho_0}{\varepsilon_0}}$$

where e and m are the electronic charge and mass, ρ_0 is the electron charge density in the beam, and ε_0 is the permittivity of free space.

In the real electron beams which occur in practical tubes the ideal plasma conditions assumed above do not hold. The beam is of finite cross-section, and radial vibrations as well as axial ones are possible. A number of modes and frequencies are possible in a real beam, but usually only the fundamental is of any significant amplitude.

The frequency of the fundamental depends to a certain extent on the geometry and the operating conditions of the tube, but it is close to—only less than—the frequency calculated for the ideal case above.

In future we shall assume that there is no significant difference between these two frequencies and we shall use ω_p to refer to the fundamental plasma frequency of a real beam, sometimes called the reduced plasma frequency.

Disturbance Imposed on Beam

Suppose we have an electron beam with a natural plasma frequency ω_p and that we impose upon it a disturbance due to a signal at a frequency ω. This might be achieved by allowing the beam to pass through an alternating electric field between two closely spaced wire meshes to which a small r.f. voltage at frequency ω is applied. This alternating field will accelerate or retard each electron as it passes between the meshes of the buncher and velocity modulation will be produced. It is convenient to consider the resultant complex electron behaviour in terms of space charge waves produced on the beam.

Each electron passing through the buncher receives some sort of impulse from the applied field (ω) and will suffer a displacement from its equilibrium position in the beam. Because it has been so disturbed it will oscillate about its equilibrium position at the natural plasma

[2] *Physical Electronics*, Hemenway *et al.*, Wiley, 1962.

frequency ω_p, while the equilibrium position travels at the beam velocity u_0. Although each electron is oscillating at frequency ω_p, the amplitude and phase of the electron displacements depend on the size and direction of the impulse received from the modulating field, i.e. it depends upon ω.

If this complex motion is analysed it is found[3] that there are two space charge waves on the electron beam which are important in the operation of amplifying valves, and all other modes are neglected.

These two waves are called the fast and the slow waves and they have phase velocities (v_f and v_s) which are typically a few per cent higher and a few per cent lower than the beam velocity.

$$v_f = \frac{u_0}{1 - \frac{\omega_p}{\omega}} \qquad v_s = \frac{u_0}{1 + \frac{\omega_p}{\omega}}$$

The phase constants of the two waves are given by $\beta = \frac{\omega \pm \omega_p}{u_0}$ and the group velocity $\left(\frac{1}{\partial\beta/\partial\omega}\right)$ is in each case u_0 the beam velocity. Thus the fast wave has a phase velocity greater than the group velocity, while the slow wave has a phase velocity less than the group velocity.

Fast and Slow Wave

The reader will be familiar with examples where complex periodic properties are analysed into a number of sinusoidal components which are considered separately. This approach is used when the example of the last section is considered in more detail.

When a disturbance at frequency ω is applied to the electron beam then a complex periodic electron axial motion is produced which can be analysed into two components: the fast wave and the slow wave, each of frequency ω but of different phase velocities.

Thus the fast wave will consist of compressions and rarefactions of the beam which travel down the tube at velocity $v_f = \frac{u_0}{1 - \frac{\omega_p}{\omega}}$. The separation between successive compressions will be $v_f \Big/ \frac{\omega}{2\pi}$.

Because these compressions and rarefactions move at a greater velocity than the beam velocity the establishment of a fast wave means that extra energy has been fed into the beam by the source producing the wave.

[3] "Wave Picture of Microwave Tubes", Pierce, *B.S.T.J.*, Vol. 33, November 1954.

The slow wave also consists of compressions and rarefactions, but these travel at v_s so that the separation between successive compressions $2\pi v_s/\omega$ is less than that in the fast wave. In this case the compressions and rarefactions travel at less than the beam velocity and, to produce a slow wave, energy must be extracted from the beam by the slow wave source.

It is worth noting that radio-frequency energy at frequency ω must be fed into a beam if the amplitude of a fast wave at that frequency is to be increased, whereas energy must be extracted if the amplitude of a slow wave is to be increased.

The unusual general properties of the fast and slow wave are well illustrated in the klystron amplifier.

Klystron Amplifier

In the simple klystron amplifier shown in Fig. III.1 an electron beam moves from left to right passing through wire meshes in the two resonant cavities at A and B. The cavities are tuned to the signal frequency ω and the alternating electric field between the two meshes in each cavity will impose a disturbance on the beam and produce waves of the type described earlier.

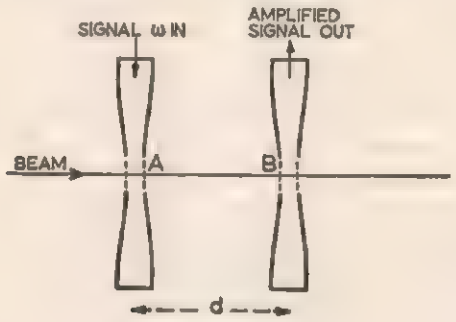

FIG. III.1

The beam to the left of the first cavity (buncher) is homogeneous and when it passes through the electric field in this first cavity each electron will receive some sort of impulse from the field and will oscillate axially at frequency ω_p. We have already seen that the phase and amplitude relation between the displacements of individual electrons is complex. The complex periodic motion may be analysed into two waves of equal amplitude: the fast wave and the slow wave, each consisting of compressions and rarefactions of the electron beam. Since the two waves are of equal amplitude the buncher feeds no nett power into the beam.

The compressions travel down the tube faster than the beam velocity u_0 in one case and slower than u_0 in the other case.

Immediately after leaving the first cavity the electrons will all have different velocities, but there will have been no time for any overtaking to occur, so the beam is still of uniform density with no bunching. Thus the compressions and rarefactions due to the fast wave exactly cancel those due to the slow wave at A, i.e. the two waves are exactly out of phase.

Because the two waves have different phase velocities, v_f and v_s, they will change phase relative to each other as they travel down the tube until at B, a distance d from A, a relative phase change of π has occurred and the two waves will be in phase, with compressions due to the slow wave adding to those of the fast wave to produce maximum bunching.

Phase constant $\beta = \dfrac{2\pi}{\lambda}$ is phase change per unit distance travelled

$$\beta = \frac{2\pi \times \text{frequency}}{\text{velocity}} = \frac{\omega}{\text{velocity}}$$

But
$$\text{velocity} = \frac{\omega u_0}{\omega \pm \omega_p}$$

$\therefore$
$$\beta_{\text{fast}} = \frac{\omega - \omega_p}{u_0} \quad \text{and} \quad \beta_{\text{slow}} = \frac{\omega + \omega_p}{u_0}$$

Relative phase difference introduced per unit distance travelled is

$$\beta_{\text{slow}} - \beta_{\text{fast}} = \frac{2\omega_p}{u_0}$$

For a phase difference of π to be introduced the distance travelled is

$$d = \frac{\pi}{2\omega_p/u_0} = \frac{\pi u_0}{2\omega_p}$$

d is the optimum bunching distance and the second cavity (catcher) is placed a distance d from the buncher.

Radio frequency oscillations will be induced in the second resonant cavity (catcher) by the bunched electron beam passing through it. The alternating electric field between the wire meshes of the cavity will impose a disturbance on the beam and generate a fast and a slow wave at frequency ω just as the buncher did.

For amplification to occur the beam must lose energy and the catcher cavity must gain it. Such a transfer of energy will occur due to interaction of the beam with the alternating electric field if the cavity is tuned so that the radio-frequency field has the correct phase.

Fast wave amplitude increases when energy is fed into the beam and slow wave amplitude increases when energy is extracted. Thus to ensure a transfer of energy from beam to catcher, the fast wave generated by the second cavity must be of such a phase that it subtracts from the original fast wave amplitude due to the buncher, while the new slow wave will add to the original one.

The beam to the right of B is still inhomogeneous but with a reduced fast wave amplitude, and an increased slow wave amplitude, the energy extracted from the beam to produce these changes appearing as radio-frequency energy in the cavity.

Travelling Wave Tube

In the klystron amplifier the two circuits, buncher and catcher, interact with the electron beam at two single points on the axis. In the travelling wave tube, and in the backward wave tube which is described later, there is a single, long, distributed radio-frequency circuit which interacts with the electron beam along its whole length. The electromagnetic wave travelling along the distributed circuit reacts with one of the waves on the beam. For this to happen the circuit wave and the selected beam wave must have approximately equal velocities.

The two beam waves will have velocities

$$v = \frac{u_0}{1 \pm \omega_p/\omega}$$

which are of the same order as the electron beam velocity u_0. This must be very much less than the velocity of light because of the difficulty of accelerating electrons to anything approaching that speed with normal H.T. voltages.

With a beam voltage V of 3000 volts the electron velocity u_0 is given by

$$Ve = \tfrac{1}{2}mu_0^2$$

$$e/m = 1{\cdot}76 \times 10^{11} \text{ coulomb/kilo.}$$

$$u_0 = \sqrt{\frac{2Ve}{m}} = (2 \times 3000 \times 1{\cdot}76 \times 10^{11})^{\frac{1}{2}} = \sqrt{10{\cdot}5} \times 10^7 \text{ m/sec}$$

Thus the beam velocity is little more than a tenth of the velocity of light and the distributed circuit will have to be one in which the velocity of electromagnetic waves is reduced to about this value.

If a waveguide is corrugated or loaded internally with discs or posts, then the wave velocity is reduced. The simplest slow wave structure is, however, the wire helix, where the wave travels down through the helix

at roughly the same speed as if it were actually travelling along the full length of the wire.

Thus if the helix has a diameter of 0·1 in. and there are 30 turns to the inch, then the wave velocity down the helix is about $\frac{c}{\pi \times 0{\cdot}1 \times 30}$, i.e. little more than a tenth of the velocity of light.

In the detailed discussion of travelling wave tube amplification we shall see that the circuit wave is required to interact with the slow electron beam wave and not with the fast one. Thus the circuit wave and the slow electron beam wave must have about equal velocities.

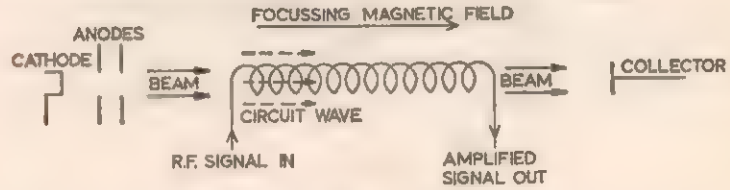

FIG. III.2

The electron beam is accordingly designed so that ω_p is large enough to give good separation between the velocities of the slow and fast electron beam waves $\left[v = \frac{u_0}{1 \pm \omega_p/\omega}\right]$. The helix is then designed to make the circuit wave velocity synchronise with the slow beam wave and not with the fast one.

In Fig. III.2 the beam produced by the electron gun is shown (full line arrows) entering and leaving the helix. To avoid confusing the diagram the arrows are not continued through the helix. The dotted arrows show the radio-frequency field of the circuit wave passing down the helix from left to right.

A steady axial magnetic field focuses the electron beam so that it has a radius comparable with that of the helix, but is not intercepted by it at any point.

When the radio-frequency signal is coupled into the left-hand end of the helix it reacts with the electron beam to produce a slow beam wave provided that the helix has been designed, and the tube current and voltage adjusted, to give synchronism between the circuit wave and the slow beam wave as described above.

The amplitude of the slow wave on the electron beam will gradually increase as it moves from left to right. But we have already seen that to increase the amplitude of the slow wave power must be extracted from the beam. In going from left to right the power in the slow wave

decreases—its negative power increases—and the positive power in the coupled circuit wave increases.

In fact, the total power contained in the coupled wave system remains constant[4] with the circuit wave growing exponentially from its original value and the beam wave "growing" similarly in the negative sense (Fig. III.3).

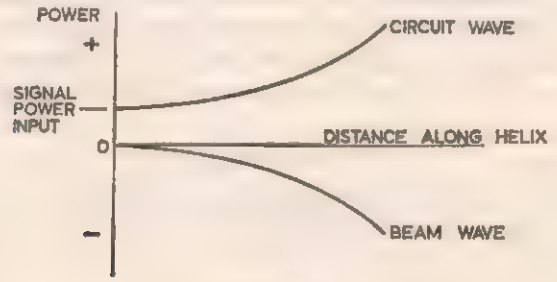

FIG. III.3

Amplification occurs along the whole length of the helix and the gain is high even although the circuit is non-resonant. The phase velocity of the wave on the helix is almost constant over a wide frequency range and the travelling wave tube has a very large bandwidth.

The great bandwidth, which might be say 1 Gc/s at 3 Gc/s, means that the tube is very useful in microwave links carrying many television or telephone channels. The original travelling wave tubes in general use fifteen years ago had noise temperatures of a few thousand degrees, but more recent techniques discussed below for limiting noise on an electron beam have given tubes with noise temperatures of a few hundred degrees.

The high gain over a large bandwidth causes problems of instability and oscillation because the coupling at the output end will certainly not be a perfect match at all frequencies. There is a possibility that the energy reflected back down the helix to the input may cause oscillation.

To avoid this oscillation, attenuation is introduced about halfway along the tube by spraying graphite over a short length of the glass which supports the helix. The currents induced in the graphite by the electromagnetic wave on the helix cause the loss.

Suppose the tube has a forward gain G in the absence of attenuation.

Let the attenuation be A.

An input signal of nominal value 1 appears at the output as G/A which is the effective amplification of the tube.

[4] "Low Noise Amplifiers for Centimetre and Shorter Wavelengths", Wade, *P.I.R.E.*, May 1961.

Suppose a fraction k of the output is reflected by a mismatch, then the signal leaving the output towards the input end is kG/A.

This will suffer attenuation and the signal arriving at the input is kG/A^2.

No oscillation will occur if $kG/A^2 < 1$.

Thus if a tube has 70 db of unattenuated forward gain, then a 40 db attenuator will certainly prevent oscillation while leaving an effective gain of 30 db.

Backward Wave Tube

Certain filter type circuits will allow the passage of waves which have their phase and group velocities in opposite directions. Such a wave is called a backward wave,[5] and if it is coupled to the slow electron beam wave then power transfer from the beam to the circuit takes place.

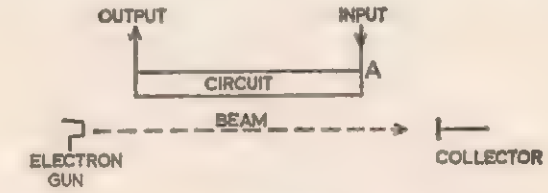

FIG. III.4

Suppose that a signal is introduced into the backward wave circuit at A (Fig. III.4), then the energy will travel down the circuit from right to left until it reaches the output end where it is removed to an external load. The phase velocity of this circuit wave is, however, from left to right, and if the power supplies are adjusted so that the slow beam wave couples with the circuit wave, then the amount of signal energy reaching the output end is greater than that entering at A. A special feature of the backward wave tube is the existence of positive feedback because the circuit wave at the output induces the slow wave on the electron beam which travels down towards the collector end of the tube where it induces a wave on the circuit at A in phase with the input signal.

Circuit waves of different frequency have different phase velocities and only one such wave will be in synchronism with the slow beam wave for a given beam velocity. Changing the beam voltage and hence the beam velocity will alter the tuning of the tube.

As described the backward wave tube could be used as an amplifier of narrow bandwidth with the input applied at A and the output taken

[5] *Travelling Wave Tubes*, Pierce, Van Nostrand, 1950.

from the circuit at the electron gun end.[6] If, however, the beam current exceeds a certain value called the starting current the tube will oscillate and it is as a backward wave oscillator that it is generally used.[7,8] There will be no input at A, and the output will be taken from the gun end as shown.

The great advantage of the backward wave oscillator is that it can be tuned over a wide range, say 7–12 Gc/s or 2·4–4·5 Gc/s, by varying the beam voltage. If the beam current is greatly increased, then oscillations at higher frequencies than the fundamental are generated simultaneously.

Noise Reduction

Most of the noise in an electron beam tube arises from shot noise produced by the cathode. This is greatly reduced if a well-designed electron gun is used. In particular, modern low noise tubes have multi-anode electron guns, with the shape and relative potentials of the anodes chosen as a result of a complex calculation which often justifies the use of analogue computation methods.[9] Commercial tubes will have a variable voltage supplied to at least one of the anodes and this is adjusted in operation to give minimum noise.

Further reduction in noise can be achieved by a sophisticated technique of stripping the residual noise power from the beam before it enters the interaction region.

We have already seen that coupling between the circuit wave and the slow beam wave results in an exponential growth of the amplitude of both waves (p. 116). The positive power in the circuit wave increases and the negative power in the slow beam wave increases, the total power in the system remaining constant.

If the circuit wave is coupled to the fast beam wave, then the total power remains constant as before but there is a gradual interchange of power between the circuit and the beam. If a particular short length of circuit is chosen, then, by the time the beam reaches the end of the circuit, all the power entering on the circuit will have been transferred to the beam, and all the fast wave power entering on the beam will have been transferred to the circuit and may be led away.

[6] "Crossed-field Backward Wave Amplifier. Noise Figure Studies", Mantena and Van Duzer, *J. Electronics and Control*, Vol. XVII, No. 5, November 1964.
[7] "Analysis of the Backward Wave Travelling Wave Tube", Heffner, *P.I.R.E.*, June 1954.
[8] "Backward Wave Oscillators", Johnson, *P.I.R.E.*, June 1955.
[9] "Analogue Method of Determining Electrode Shapes of Electron Guns", Lomax, *J. Electronics and Control*, Vol. XV, No. 3, September 1963.

The electron beam may be passed near such a short length of circuit immediately after leaving the gun and before it enters the main interaction region. The beam will be carrying both slow and fast wave noise power on leaving the gun. The signal is applied to the start of the short length of circuit. At the end of the circuit all the signal power will be on the fast beam wave and all the fast noise power will have been removed from the beam into the circuit and led away.

We now have a noise-free fast wave carrying signal power ready to be amplified, but we also have a slow wave carrying noise. If the beam velocity is now suddenly increased the slow wave will be eliminated[10] leaving a noise-free fast wave carrying signal power to be amplified.

The amplification process cannot, however, be the ordinary travelling wave tube technique described earlier because this uses a slow beam wave and we have launched a noise-free fast wave. Amplification using a fast wave can be achieved by supplying power at twice the signal frequency. Beam tubes using this principle are called electron beam parametric amplifiers and are discussed elsewhere (p. 81).

Practical Electron Beam Tubes

Descriptions of the structure of practical electron beam tubes are available in such references as [11,12,13], and manufacturers' data sheets give performance figures. Such literature will contain details of beam voltage and current, power output, gain, noise figure, and tuning range, focusing magnetic field and whether it is provided by a solenoid or permanent magnet system.[14] Waves, usually of fairly low frequency, associated with the heavy gas ions in a plasma can have analogous behaviour to the electron beam waves already described, and some of their properties are outlined in reference [10].

In general low noise, wide-band, low power travelling wave tubes are now available with noise temperatures of about 500 degrees absolute, and tubes with higher noise temperatures may give many watts output with large bandwidth.[15] For very high power output multicavity klystrons several feet tall have high efficiency and give many kilowatts of mean power.

[10] "Role of Space Charge Waves in Modern Microwave Devices", Sims and Stephenson, *Electronic Engineering*, July 1960.
[11] *Microwave Engineering*, Harvey, Academic Press, 1963.
[12] "Travelling Wave Tubes for Communications Networks", Coulson, *Marconi Point to Point Telecomms.*, Vol. 4, No. 2, February 1960.
[13] "Backward Wave Oscillator for Millimetre Waves", King, *Marconi Review*, Vol. XXV, No. 146, 1962.
[14] "Compensated Reversed Field Focussing of Electron Beams", Burke, *P.I.E.E.E.*, Vol. 51, No. 11, November 1963.
[15] *Power Travelling Wave Tubes*, Gittins, E.U.P. (1965).

Crossed Field Tubes[16]

In the ordinary (or *O*-type) electron beam tubes discussed above the power required for the amplification process is drawn from the kinetic energy of the electrons. These electrons must be kept in approximate synchronism with the R.F. wave travelling on the helix or other wave-supporting structure. Consequently there will be considerable energy left in the beam when it has ceased to interact with the wave, and *O*-type tubes will be of low efficiency.

If electrons are injected into a region where there are crossed electric and magnetic fields, then the electron paths become complex, but the drift velocity of the electrons through the region, which also contains the wave-supporting structure, remains fixed and is independent of the electron velocity.

By a suitable choice of the crossed electric and magnetic fields the electron drift velocity can be made to synchronise with the structure-borne R.F. wave. Even although this wave receives much energy from the electrons (which in turn receive it from the electric field), synchronism will still be maintained, regardless of the length of the interaction region.

Amplifiers and oscillators using crossed fields in this way are called *M*-type tubes. They may be forward or backward wave operating and are very much more efficient than *O*-type tubes and are capable of very high power operation. *M*-type tubes have, however, poor noise qualities and, while they deserve mention in a general appendix of this sort on electron beam tubes, they have no low noise application.

Although *M*-type operation can be used with a long tube of the conventional travelling wave type, it is much more convenient to bend the wave-supporting structure into a circle and have a more compact magnetic field. The most familiar example of the *M*-type tube is probably the magnetron, a forward wave oscillator, which was invented before *M*-type operation was fully understood.

If the magnetron anode structure, which supports the R.F. wave, is broken to allow an input and an output, then the valve may be used as a backward wave amplifier (amplitron). If an external frequency-determining feedback circuit is connected between output and input then a backward wave oscillator (stabilotron) is obtained. Both these tubes are increasingly used in commercial and military systems.

The amplitron allows a complex signal, such as the pulse of a sophisticated radar system, to be generated at low level and then amplified with good efficiency to a high power level. The stabilotron is

[16] *Crossed Field Microwave Devices*, Okress, Academic Press (1961).

tending to replace the magnetron because of its greater frequency stability and efficiency.

Another *M*-type backward wave oscillator is the carcinotron which has a localised cathode rather than the extended cylindrical cathode of the magnetron and stabilotron. Very broad band tunable operation at high power is possible with the carcinotron which is likely to be principally used in electronic warfare as a jammer, where the noisy character of the output is little disadvantage.

Appendix IV

Ferrites and their Microwave Applications

THE FERRITES AVAILABLE SINCE 1946, are a special group of magnetic materials with very high resistivity. Because of their high resistance, eddy current effects are much reduced and ferrites can be used as cores for transformers and coils in the high frequency range where conventional metal cores in laminated or dust form become inefficient.

The origin of the magnetism in strongly magnetic materials is the spinning electron. If an electromagnetic wave is propagated through a magnetic medium, then there is interaction between the spinning electrons and the alternating magnetic field of the propagated wave. This interaction is strongest at frequencies in the microwave region but in metals eddy currents cause the skin depth to be very small. The high resistance of the ferrites, however, means that the alternating magnetic field will penetrate them in depth, even at ultra-high frequencies, so that the effects of this interaction between a microwave signal and the electron spin can be readily observed and used.

It can be arranged that the interaction between the microwave signal and the electron spin in the ferrite results in such effects as the absorption of energy from the signal or in the introduction of a phase change.

The character and amount of the interaction between signal and spin is affected if a steady magnetic field is also applied to the ferrite. Variation of this steady magnetic field can thus be made the means of producing a variable attenuator or a variable phase changer in a waveguide system.

The nature of the interaction between the electron spin and the microwave signal depends upon the relative directions of the polarisation of the signal and the steady applied magnetic field. For a given microwave signal, propagation through a piece of ferrite with a steady applied magnetic field will produce one effect for propagation in one direction, and a different effect for propagation in the opposite direction.

Such unidirectional effects enable the microwave isolators, circulators, etc., which are described below, to be realised.

In Fig. IV.1 the isolator represents an element which gives complete attenuation for signals passing from left to right, but zero attenuation for similar signals passing from right to left.

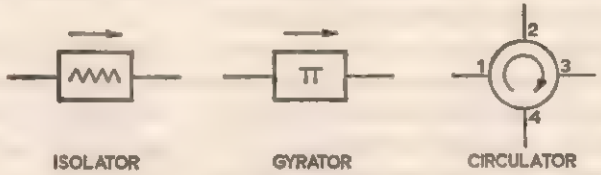

FIG. IV.1

In the gyrator a phase change of π radians is introduced into signals going from left to right, but no phase change for signals going from right to left.

In the four-port circulator signal can only pass from any one port to its next neighbour in the clockwise direction. Passage from 1–2 or 2–3 or 3–4 or 4–1 is possible, but infinite attenuation is encountered in any other path such as 1–4 or 3–1.

Structure, Resistance, and Magnetism

The ferrite molecule has the general formula MFe_2O_4 where M is some divalent metal such as nickel, cobalt, or magnesium, and the two iron atoms are trivalent ferric atoms. In the naturally occurring ferrite Fe_3O_4 (magnetite or ferrous ferrite) M is a divalent ferrous atom.

A finely powdered intimate mixture of metal carbonates or nitrates is prepared so that the correct proportions of M, Fe, and O atoms will appear in the ferrite molecule after chemical combination. The chemical combination and sintering may be carried out at the same time by heating the powder to a high temperature under pressure. Crystal growth takes place and a dense uniform mass of ferrite is produced. The material is hard and, to avoid excessive working, it is often prepared in moulds to give roughly the shape finally required.

Since the metal atoms in the material are in strong chemical combination in the ferrite molecule, the electrons are more tightly bound than in pure metals and the resistance of the ferrites is much greater than that of the magnetic metals. Typically, the resistivity of a ferrite is many million times that of iron.

The crystal structure of the ferrite consists of a face-centred cubic structure of oxygen ions with the metal ions M and Fe fitted in between.

There are two possible types of site for these metal ions and the two ferric ions lie on different types of site. Thus, for example, in a nickel ferrite where M is nickel, one ferric ion and the nickel ion are on one type of site, and the other ferric ion is on the other type of site. The siting of the metal ions is important in establishing the difference between the origins of the magnetism in the pure metals (ferromagnetism) and in the ferrites (ferrimagnetism).

In the ferromagnetic metals the magnetic moments of all the atoms in any one region or domain are aligned in the same direction. In the unmagnetised state the magnetic moments of adjacent domains are orientated at random, so that there is no nett magnetisation. An applied steady magnetic field causes whole domains to swing round so that their magnetic moments lie in the direction of the field.

In the ferrimagnetic material the magnetic moments of some of the atoms in the molecule oppose each other. In particular, in nickel ferrite the magnetic moments of the atoms on different types of site tend to oppose each other. Thus the resultant magnetic moment per molecule is not that due to the arithmetic sum of two ferric atoms and a nickel atom, but is much closer to that of the nickel atom alone. It is the resultant magnetic moments of the ferrite molecules which are aligned in the same direction in any one domain.

Gyromagnetic Effects

A spinning electron possesses a magnetic moment and an angular momentum which are in opposite directions. If a steady magnetic field H_0 is applied to a magnetic material the direction of the magnetic moment of the spinning electron tends to align itself with the field. However, due to the gyroscopic effect of the associated angular momentum, precession takes place and the end of the magnetic moment vector describes a circle around the direction of the field H_0 in a plane perpendicular to that direction. Due to damping forces in the material the amplitude of the precession gradually decreases, i.e. the circle gets smaller, until the magnetic moment of the spinning electron and the direction of the field are aligned. The material is then magnetised. The frequency of the precession depends on the strength of H_0.

If the material is placed in a steady field H_0 so that the spin magnetic moments are already aligned, then the application of an alternating magnetic field with a component perpendicular to H_0 will disturb this alignment and precession will occur. Interaction takes place between the precessing magnetic moments of the spinning electrons and the alternating magnetic field. This interaction produces phase changes, attenuation, etc., in the alternating field, and these effects are used in the

microwave devices described below. The interaction is dependent on the frequency of the alternating field for a given H_0, and upon the polarisation of the alternating field relative to the direction of H_0.

The state of polarisation of the alternating signal applied to the ferrite can invariably be described in terms of circular polarisation. Thus the commonly encountered plane polarised wave is equivalent to two circularly polarised waves of equal amplitude but of opposite senses of polarisation, i.e. one clockwise and one anti-clockwise.

In a magnetised ferrite there are two different phase velocities for the two different components of the applied alternating signal. That component of the signal which is circularly polarised in a clockwise sense about the direction of the steady field has one phase constant (β_+), and the component which is circularly polarised in the opposite sense is propagated with a different phase constant (β_-) *in any direction in the material.*

After passing through a thickness d of magnetised ferrite two waves circularly polarised in opposite senses about the steady field will each undergo different phase changes and, if originally in phase, will emerge with a phase difference $(\beta_+ - \beta_-)d$ between them.

In addition to this differential phase change for the two senses of circular polarisation, there may be differential absorption.

For a given steady field H_0 there is a certain precession frequency for the spinning electron. If an alternating signal of this frequency and circularly polarised in a positive sense is applied to the ferrite, then resonance occurs and a large amplitude of precession is excited. There is consequently a large absorption of energy from the signal due to the forces damping the precession. Had the signal been polarised in the opposite sense, then this resonance absorption would not have occurred.

Thus the differential behaviour of magnetised ferrites is dependent upon the sense of the circular polarisation of the applied microwave signal relative to the steady field. There are two such differential phenomena: in the first case the phase constant of the positively circular polarised wave differs from that of the negatively polarised wave; while in the second case resonance absorption occurs for the positively polarised wave but not for negative polarisation.

For ferromagnetic resonance at microwave frequencies high values of H_0 are needed, and devices depending on resonance absorption are sometimes called strong field devices. If the other differential effect depending on phase constant is to be employed, then low values of H_0 are used so that there is no possibility of resonance absorption introducing unwanted effects.

An appreciation of gyromagnetic behaviour and its applications is

probably best obtained by considering the individual microwave devices described below.

Phase Shifter for Circularly Polarised Waves

A ferrite rod is mounted in polyfoam along the axis of a circular waveguide. A magnetic field H_0 is directed along the length of the rod by a solenoid carrying current. If a wave of positive circular polarisation is passed along the guide, then it will undergo a phase change which will depend upon the length of ferrite traversed and upon the field H_0. The phase change can be varied by varying the current through the solenoid. A different phase change would be obtained if the same wave were sent through the guide in the opposite direction because it would now be negatively circularly polarised about the field direction.

Rectangular waveguide systems carrying plane polarised waves are in practice more common than circular guides. The variable phase shifter described above can, however, be used in a rectangular system by inserting two quarter wave plates mutually perpendicular, one before and one after the ferrite rod.

A quarter wave plate may consist of an insulating plate, or a set of vanes, to introduce a phase change of 90° between components perpendicular and parallel to the plate. Thus a plane polarised wave is converted to a circularly polarised wave if the quarter wave plate is set at 45° to the plane of polarisation of the plane polarised wave.

The first quarter wave plate converts the plane polarised wave to circular polarisation so that it may be operated on by the ferrite rod. The circularly polarised wave, delayed in phase by a given amount after passing through the rod, is now reconverted to a plane polarised wave by the second quarter wave plate.

Faraday Rotation—Isolator—Gyrator

It can be shown that a plane polarised wave is equivalent to two circularly polarised waves of equal amplitude but opposite sense. These two circularly polarised waves will be propagated through a longitudinally magnetised ferrite with different phase constants, so that when they are compounded after passing through the ferrite, the direction of plane polarisation of the resultant is different from that incident on the rod. The plane of polarisation has been rotated by an amount depending upon the longitudinal magnetic field and the length of ferrite. This phenomenon is known as Faraday rotation, and in conjunction with absorbing elements set at the correct angle, or suitably twisted waveguide, it is used as the basis of isolators or gyrators.

A resistive card in a waveguide will strongly attenuate a wave passing

over it with its electric vector polarised in the plane of the card, while a wave polarised perpendicularly to the card will hardly be affected.

Two such cards are placed, one at each end of a longitudinally magnetised ferrite rod mounted in a circular guide so that a 45° clockwise Faraday rotation is obtained for signals going from left to right. The resistive card on the right is at 45° to the other card, also in the clockwise direction viewed from the left. A signal entering the system from the left and passing unattenuated over the left-hand card is still polarised perpendicular to the plane of the right-hand card and is unattenuated.

A similar signal entering the system from the right and passing unattenuated over the right-hand card will again be rotated through 45° in the same direction as before—clockwise about the direction of the *field*—so that now it meets the left-hand card polarised in the plane of the card and is highly attenuated.

Circular-to-rectangular tapers are provided to allow the isolator to be used in a normal rectangular waveguide system. The rectangular tapers at each end of the isolator described above will be twisted 45° relative to each other just like the resistive cards.

In the gyrator a 90° twist is put in the taper on one side of a circular guide containing a ferrite rod giving 90° clockwise rotation about a steady magnetic field directed from left to right. For a signal going from left to right, say, the 90° twist is added to the 90° rotation giving 180° rotation, which is equivalent to a phase change of 180°. For a signal going the opposite way the 90° twist is subtracted from the 90° rotation giving no resultant effect.

Circulator

A four-port circulator based on Faraday rotation is produced by having four rectangular guide inserts into the circular guide. The guide contains a magnetised ferrite rod which gives a 45° rotation. The two inserts on the left of the rod (ports 1 and 3) are at right angles to each other, so that no signal will pass between them. The two inserts on the right of the rod (ports 2 and 4) are also mutually perpendicular and twisted 45° relative to 1 and 3. The senses of the Faraday rotation and the twist are such that there is no nett rotation in going from 1 to 2, but 90° in going from 2 to 1 and hence no coupling. Similarly, signal will pass from 2 to 3, 3 to 4, and 4 to 1 but not in the opposite sense.

Transverse Fields—Rectangular Guides

The devices so far described all depend upon a longitudinally magnetised ferrite rod in a circular waveguide which is fitted into a standard

rectangular system by suitable tapers. It is, however, possible to get gyrator and isolator effects directly in rectangular guides by putting a ferrite slab near and parallel to one of the short sides of the rectangular section, with H_0 applied perpendicular to the long sides.

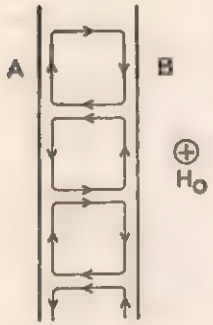

FIG. IV.2

The instantaneous magnetic field in a rectangular guide is shown in Fig. IV.2, the loops moving in the direction of propagation. Let us assume that the wave is propagated up the page. Then a point on the side A will experience a magnetic field vector rotation anticlockwise in the plane of the paper, while on the side B the vector will rotate in the opposite sense about the steady transverse field H_0 directed into the paper. If the direction of propagation is now reversed, then the sense of the rotation at A and B will also be reversed. Thus if a ferrite slab is placed near the side A, then there will be different phase velocities for waves propagated in different directions and hence a unidirectional nett phase shift.

If the field H_0 is increased until ferromagnetic resonance occurs, then the unidirectional gyrator described above becomes an isolator.

Other Ferrite Devices

The microwave devices discussed briefly above represent only a few of the applications of the special properties of ferrites. Other properties associated with gyromagnetism, like the concentration of negatively polarised energy in the ferrite, are used to give unidirectional microwave effects. The low frequency uses as cores have been mentioned in the introduction, and there is a whole important field of application in computing based on the square hysteresis loop which can be obtained in certain ferrites.

The references quoted below give details of most applications.[1,2]

[1] *High Frequency Applications of Ferrites*, Roberts, E.U.P.
[2] *Proc. I.E.E.*, Vol. 104, Part B, Supplements 5, 6, 7.

Appendix V

The Injection Laser

LASER ACTION IN A p–n junction was first obtained in November 1962, using gallium arsenide.

The importance of the injection laser lies in its small size and simplicity and the fact that it can be modulated directly by varying the stimulating current flowing across the p–n junction. In other lasers the output is modulated by fairly complicated electro-optical devices operating on the light beam itself. The first successful transmission of voice signals over an injection laser beam was made in the early summer of 1963.

Two Level Maser Action

In any system which shows two level maser action the individual units of the system have two permitted energy levels E_2 and E_1, where

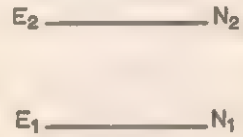

FIG. V.1

$E_2 > E_1$. In a normal system in equilibrium at a given temperature the population in level E_2 is much smaller than that in E_1.

$$N_2 = N_1 e^{\frac{-(E_2-E_1)}{kT}}$$

The system is intended to amplify signals of frequency f, where $E_2 - E_1 = hf$.

A photon of such radiation entering the system may be absorbed by a unit in state E_1, which will thereby be excited to state E_2. Alternatively the incident photon may stimulate one of the units in state E_2 to fall to

state E_1, emitting as it does so a photon of the same frequency as the incident photon and coherent with it.

There may also be an occasional transition of a unit from E_2 to E_1 unconnected with the arrival of an incident photon. Such spontaneous emission, which is not coherent with the signal, constitutes noise and will not be considered further here.

A photon of energy in the signal incident on the system may thus be absorbed if it meets a unit in state E_1 or amplified if it meets a unit in state E_2. But since the number (N_2) of units in state E_2 is much less than the number (N_1) in state E_1 there is a much greater probability of absorption than of stimulated emission. A signal incident on a normal system in temperature equilibrium with its surroundings will be attenuated not amplified.

If amplification is to be achieved in the material, then a non-equilibrium population distribution must be achieved and maintained, i.e. N_2 must be increased relative to N_1. This non-equilibrium state is usually obtained by pumping, i.e. by supplying energy from a power source to pump units from state E_1 to state E_2. Thus N_1 is decreased and N_2 is increased, and the signal power is amplified at the expense of power from the pump.

The required increase in N_2 relative to N_1 could in principle have been achieved just as well not by pumping some of the N_1 population up to N_2, but by introducing into the system from outside enough extra units already in the higher energy state E_2. This process is used in the injection laser.

Simple Injection Mechanism

Suppose that a material can exist in the n and p forms due to the presence of impurity centres giving the levels shown in Fig. V.2 (i).

If an abrupt discontinuity between the n and p regions can be arranged in such a material, then a rectifying junction will be formed (p. 26).

Such a junction biased in the forward direction by an external battery is shown in (ii) and we shall concentrate on that portion of the forward current which consists of electrons injected into the p region from the n region. This will be the predominant part of the current if the conductivity of the n-type material is made much greater than that of the p-type. The most energetic of the electrons in the conduction band of the n-type material, those shown shaded on the distribution curve, will have sufficient energy to pass into the p-type material.

The p-type material thus has electrons injected into it at a high energy level. In terms of the two level discussion above, the population N_2 at

the higher level E_2 is increased relative to the low level population N_1.

A photon of frequency, $f = \dfrac{E_2 - E_1}{h}$, travelling in the p-type material will thus have an increased chance of stimulating emission of another, coherent, photon by causing an electron to fall from E_2 to E_1. Note that it is possible for electrons to fall into E_1 because the material is p-type with holes in the valence band.

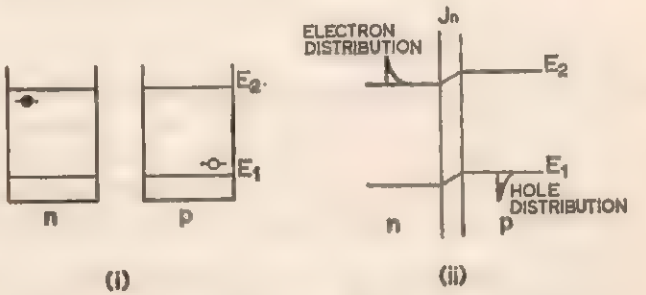

FIG. V.2

The photon travelling in the p-type material might have been absorbed in exciting an electron from E_1 to E_2, but the chance of such absorption occurring will decrease relative to the chance of stimulated emission as the number of electrons injected into E_2 across the junction increases, i.e. as N_2 increases relative to N_1.

If the forward current is sufficiently increased there will be a nett gain of light energy in the p-type material. If two silvered optical flats are ground on to the light emitting layer of the diode then the coherent photons which are perpendicular to the faces are reflected to and fro and are amplified by the mechanism described above. If one of the faces is not made fully reflecting, then sufficient energy is fed back to maintain the optical resonator in oscillation, while the remainder is emitted as a coherent beam.

Practical Injection Junction Materials

A population inversion of the type described above in a p-type material could equally well have been produced by injecting holes as high energy minority carriers into an n-type semiconductor. It should be remembered that if a hole is given extra energy it goes lower in the energy level diagram. Thus the injection of holes from a p-type material across the junction into the valence band of an n-type material produces the required non-equilibrium distribution, and in general the establishment of an excess population of high energy minority carriers in either p- or

n-type material will produce transitions across the energy gap giving photon amplification and the possibility of laser action.

The energy gap in germanium is 0·72 eV, so that the energy involved in a transition is $0{\cdot}72 \times 1{\cdot}6 \times 10^{-19}$ joule.

Thus $hf = 0{\cdot}72 \times 1{\cdot}6 \times 10^{-19}$ joule.

and $$\lambda = \frac{ch}{0{\cdot}72 \times 1{\cdot}6 \times 10^{-19}} = \frac{3 \times 10^{8} \times 6{\cdot}6 \times 10^{-34}}{0{\cdot}72 \times 1{\cdot}6 \times 10^{-19}} m$$

$$= 17{,}200 \text{ A.U.}$$

The silicon energy gap of 1·09 eV gives a larger transition energy, but the wavelength involved is still very long, and transitions in both germanium and silicon result in heating of the material and no light output. Gallium arsenide has a gap of 1·4 eV and here a direct transition gives a wavelength of about 8000 A.U., so that an infra-red beam will be emitted.

If visible light is required, then a material with a larger transition energy must be used. Such a material is gallium phosphide where the energy gap is 2·25 eV.

In a transition process momentum must be preserved as well as energy, and with the increase in energy gap direct transitions become less probable. A transition takes place via an impurity centre which can absorb or emit phonons thus simplifying the conservation of energy and momentum conditions. Thus an electron does not cross the gap directly but goes via an impurity centre (p. 38).

Suitable impurity must be introduced into the gallium phosphide, so that the transition process can occur with the radiation of visible light. Conversely, one of the important requirements in preparing the gallium phosphide is that other impurity centres which would give rise to non-radiative transitions are reduced to a minimum.

Different impurities give luminescent centres at different depths in the energy gap. Doping gallium phosphide with oxygen and zinc produces green light and must correspond to a shallow impurity centre near the band edge. Doping with oxygen and silicon gives red light, the longer wavelength indicating that the impurity centre is in this case deeper; it is in fact thought to be about 0·4 eV from a band edge.

Silicon carbide, which was the material in which Lossev first observed electroluminescence, has a large energy gap, but is difficult to manufacture. In general, the greater the energy gap the more difficult is manufacture and attention is directed to solving these problems with promising new materials and to discovering more about the desirable luminescent centres and the undesired non-radiative centres.

Injection laser action has been observed in a large number of other materials, e.g. gallium arsenide, indium antimonide, and mixed crystals of indium gallium arsenide. By choosing a given crystal material, or by selecting a certain proportion of elements in a mixed crystal, a wide choice of wavelengths is available. In a given material the operating wavelength may be varied slightly by an applied magnetic field[1] and pressure tuning is also possible.

New materials, including group IV–VI compounds are extending laser action into the long-wave infra-red where atmospheric "windows" allow useful transmission, and new pumping and tuning methods are likely to be used in commercial and military applications in the near future.[2]

Limitations of Injection Lasers[3]

The conversion efficiency (electricity to light) falls as the temperature rises and the current required for laser action increases. The junction laser can, however, be continuously operated at temperatures up to about −77° C.

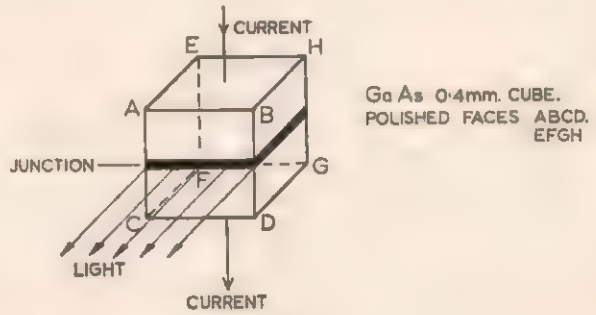

FIG. V.3

The frequency response is limited by the junction capacity and its series resistance, but the upper limit is probably several hundred Mc/s. The spectral width of the emitted light may however be several A.U. which is much greater than that of the ruby or gas laser, but careful control of operating conditions can considerably reduce this.

[1] "Semiconductor Lasers", Quist, *International Science and Technology*, February 1964.
[2] "Semiconductor Lasers", Nathan, *P.I.E.E.E.*, Vol. 54, October 1966.
[3] "Considerations on the Applications of Semiconductor Lasers", Milsum, *I.E.E. Laser Symposium*, September 1964.

A publicised feature of the laser is the extreme narrowness of the light beam which is determined only by diffraction from the emitting area. The active area of a single junction laser (Fig. V.3) is, however, so small that diffraction gives a beamwidth of several degrees rather than the fraction of a degree which may be obtained with other types of laser beam.

Appendix VI

The Requirements for Low Noise Amplification Near 1 Gc/s

THE CONCENTRATION OF THE new types of low noise amplifying systems in radio-astronomy on frequencies in the general vicinity of 1 Gc/s is a combination of accident and design.

A highly directional system must operate in the many megacycles per second region if the aerial is to be of reasonable size, and so too must any system requiring signals to penetrate the ionosphere. Furthermore, background noise from the sky falls to a minimum at about 1 Gc/s.

Unfortunately the excess noise introduced by U.H.F. triode and transistor amplifiers increases very sharply between 100 Mc/s and 1 Gc/s, while crystal mixers also have high noise figures in this frequency range. Thus any conventional amplifier introduces noise which is several orders of magnitude greater than the natural background.

Typically at 1 Gc/s the noise temperature of a conventional radio receiving system might be about 2000° K, while the sky temperature is only about 10° K. If the system noise could be reduced to about the same order as the sky noise, then the sensitivity would be enormously increased.

Such an improvement is possible if a parametric amplifier or a maser is used in the receiver. Parametric amplifiers have a high frequency limit of several Gc/s and masers a low frequency limit of about 1 Gc/s. Between them these devices cover the whole frequency range where the sky noise is low (about 0·5–10 Gc/s). In particular, either can be used at 1420 Mc/s, the hydrogen line frequency (see p. 138), which is of vital interest in radio astronomy.

System Noise

System noise is made up of the following components: aerial overspill, aerial and feeder losses, device noise, and second stage noise.

Aerial Overspill. Back lobes and side lobes of the directional aerial system may be pointed at noise sources, e.g. the ground, at relatively high temperatures like 300° K.

Aerial and Feeder Losses. Any attenuating element introduces into the system an amount of noise dependent on the attenuation and the temperature of the aerial, switch, waveguide, etc., producing the attenuation (see p. 9).

Device Noise. The noise introduced by whatever device is used for the first stage of amplification is the principal component of the noise produced in the receiver itself. Typical noise temperatures of some conventional devices used at these frequencies are:

Triode valve 120° K at 100 Mc/s, 1200° K at 1 Gc/s.
Crystal mixer 500–1000° K.
Transistor 1500–2000° K.
Tunnel diode amplifier 1000° K.

Second Stage Noise. In a low noise system the noise introduced by the second stage may be significant. Thus a conventional second stage with a noise temperature of 1000° K, following a first stage which has a power gain of 100, will add 10° K to the system noise temperature.

All these factors must be considered if a system is to be produced with the lowest noise temperature. The substitution of a low noise amplifier for the conventional first stage will make it worth while to redesign the associated aerial, feeders, etc. for low noise properties.

Three stages in the improvement of the total noise temperature of a system might be:

Conventional amplifier with conventional aerial, feeders, etc., 2000° K.

Low noise amplifier with conventional aerial, feeders, etc., 100–200° K.

High quality specially designed maser system (see p. 9), 18·5° K. Practical limit probably a few degrees absolute.

Sky Noise

The two components of sky noise are galactic noise and atmospheric noise.

The origin of galactic noise is uncertain, but it consists of a continuous background whose intensity varies with celestial direction but not with time. Superimposed on this background are various discrete

sources or radio stars. The background noise falls sharply with frequency, in the manner shown in Fig. VI.1, so that the noise temperature is about 100° K at 100 Mc/s but only a few degrees absolute at 1 Gc/s.

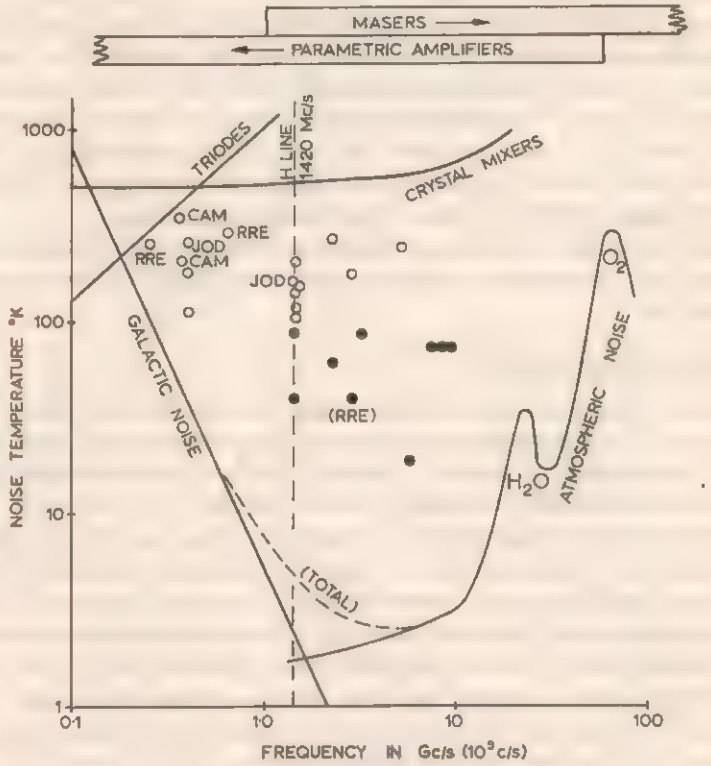

FIG. VI.1

Energy is absorbed when it passes through the Earth's atmosphere and the attenuation results in noise as shown in Chapter 1, p. 9. The mechanism by which oxygen and water molecules absorb energy has been discussed in Chapter 2, p. 14.

Figure VI.1 shows the way in which atmospheric noise increases with

frequency, and the position of the noise peaks. The minima or "windows" above 100 Gc/s are of interest in millimetric systems, but the attenuation in this region due to water vapour appears higher than expected, particularly near oxygen lines.[1]

Figure VI.1 also shows galactic noise becoming negligible at frequencies above a few Gc/s, and atmospheric noise becoming negligible at frequencies below 1 Gc/s. Thus the total sky temperature has a minimum value at a few Gc/s, and a generally low value between, say, 0·5 and 10 Gc/s. The indicated operating bands of masers and parametric amplifiers both extend into this low sky temperature region and the two bands overlap. Either device may thus be used at 1420 Mc/s, the frequency of the hyperfine hydrogen line (see below).

Also shown on the diagram (dots) are the noise temperatures of current low noise systems (see below) and of conventional triodes and crystal mixers.

Hydrogen 1420 Mc/s Line

Hyperfine structure in spectra is caused by nuclear spin. With the nucleus of the hydrogen atom spinning in a given direction there are two possible orientations for the spinning electron, each corresponding to a given atomic state and energy. Transitions from one of these states to the other involve the emission or absorption of a quantum of energy. The frequency of this radiation is 1420 Mc/s (compare caesium where the single outermost electron reacts with the magnetic moment of the spinning nucleus to give the 9192 Mc/s of the caesium atomic clock).

The probability of a hydrogen 1420 Mc/s transition occurring is very low—about one transition per atom in ten million years—but the amount of atomic hydrogen in space is so great that it was correctly predicted in 1944 that the 1420 Mc/s radiation would be detectable.

About 90% of the interstellar gas is atomic hydrogen, with most of the remainder being ionised hydrogen in the vicinity of the hot stars. Radiation from the atomic hydrogen is able to penetrate the dust which obscures the centre of the galaxy when viewed with the great optical telescopes.

The concentration of gas at the galactic centre and the spiral form of the arms can be observed, and the motion of the gas has been established by Doppler frequency shift measurements.

The existence of the 1420 Mc/s hydrogen line and the fact that it occurs in the sky temperature minimum has been of tremendous importance to the astronomers and cosmologists.[2]

[1] Tolbert and Straiton, *Proc. I.R.E.*, March 1961, p. 649.
[2] *The Exploration of Outer Space*, Lovell, O.U.P.

Existing Low Noise Systems

The dots in Fig. VI.1 indicate the total noise temperature of some existing low noise installations in the radio astronomy field. Some of these systems employ pencil beams through which the signal source drifts due to the Earth's rotation, some track the source the whole of the time it is above the horizon. Other systems use interferometer techniques, with two spaced aerials in the simplest case and with multiple aerials and elaborate phase switching in the more complex high resolution instruments.[3,4]

These interferometer techniques distinguish between small and extended sources and thus tend to reduce the effect of the galactic noise background. Thus even at the lower frequencies where galactic noise is high it is worth using a low noise amplifier. Parametric amplifiers are used in this low frequency application.

In any case the observation or "integration" time needed to detect a given signal increases with the system noise and this is an important consideration in radio astronomy where a long integration time may entail tracking the source with a heavy aerial system.

In general, the maser systems operate at higher frequencies and have lower noise figures than the parametric systems, although they do overlap in frequency and occasionally in noise figure. The parametric amplifier, usually employing a solid state diode or an electron beam, is a simpler device to operate than the maser. In particular, it does not need helium cooling or a magnet, so that it is smaller and lighter to fit on the aerial system. The operational balance may be restored by the development of compact masers with closed cycle helium liquefiers and superconducting magnets.

Outside the "Window"

The requirement for low noise amplification arises in earth-bound receivers studying radiation from space because of the radio "window" at about 1 Gc/s. If, however, the receiver is taken outside the atmosphere in a satellite or even up in a high flying balloon, then higher frequencies into the millimetre and infra-red regions can be detected. At these very high frequencies compact masers or maser-type photon counters are likely to be increasingly employed.

[3] Ryle, *Proc. Roy. Soc.*, Vol. *211A*, 1952, p. 351.
[4] Christiansen and Mathewson, *Proc. I.R.E.*, Vol. 46, 1958, p. 127.

Bibliography

General Books

Wave Mechanics of Crystalline Solids, Smith, Chapman & Hall.
Physical Electronics, Hemenway *et al.*, Wiley.
Microwave Tubes and Semiconductor Devices, Sims and Stephenson, Blackie.

Specialised Books

Coupled Mode and Parametric Electronics, Louisell, Wiley.
Varactor Applications, Penfield and Rafuse, M.I.T. Press.
Power Travelling-wave Tubes, Gittins, E.U.P.
Lasers, Lengyel, Wiley.
Applied Optics. "Supplement on Optical Masers", 1962.
P.I.E.E.E./Applied Optics. "Joint Issue on Optical Electronics", October 1966.

Recommended Journals

International Science and Technology, New York.
Proceedings of Institute of Electrical and Electronic Engineers; also associated publications, *I.E.E.E. Spectrum and I.E.E.E. Transactions on Electronic Devices*.
Journal of Institution of Electrical Engineers.

Index

Index

Books in the same series

HIGH FREQUENCY COMMUNICATIONS
J. A. Betts, B.Sc., Ph.D., C.Eng., M.I.E.E.

In preparation

HYBRID COMPUTATION
CRYOELECTRONICS

Low Noise Electronics

W.P.JOLLY